

EMPIRICAL RESEARCH AT A DISTANCE: NEW METHODS FOR DEVELOPMENTAL SCIENCE

EDITED BY: Dima Amso, Rhodri Cusack, Lisa Oakes, Sho Tsuji and
Natasha Kirkham

PUBLISHED IN: Frontiers in Psychology





frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-88976-382-5

DOI 10.3389/978-2-88976-382-5

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

EMPIRICAL RESEARCH AT A DISTANCE: NEW METHODS FOR DEVELOPMENTAL SCIENCE

Topic Editors:

Dima Amso, Brown University, United States

Rhodri Cusack, Trinity College Institute of Neuroscience, Ireland

Lisa Oakes, University of California, Davis, United States

Sho Tsuji, The University of Tokyo, Japan

Natasha Kirkham, University of London, United Kingdom

Citation: Amso, D., Cusack, R., Oakes, L., Tsuji, S., Kirkham, N., eds. (2022).
Empirical Research at a Distance: New Methods for Developmental Science.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-382-5

Table of Contents

- 06 Editorial: Empirical Research at a Distance: New Methods for Developmental Science**
Sho Tsuji, Dima Amso, Rhodri Cusack, Natasha Kirkham and Lisa M. Oakes
- 11 From Lab to Zoom: Adapting Training Study Methodologies to Remote Conditions**
Valerie P. Bambha and Marianella Casasola
- 24 Measuring Emerging Number Knowledge in Toddlers**
Alex M. Silver, Leanne Elliott, Emily J. Braham, Heather J. Bachman, Elizabeth Votruba-Drzal, Catherine S. Tamis-LeMonda, Natasha Cabrera and Melissa E. Libertus
- 36 Studying Children's Eating at Home: Using Synchronous Videoconference Sessions to Adapt to COVID-19 and Beyond**
Shruthi Venkatesh and Jasmine M. DeJesus
- 50 Research at a Distance: Replicating Semantic Differentiation Effects Using Remote Data Collection With Children Participants**
Catarina Vales, Christine Wu, Jennifer Torrance, Heather Shannon, Sarah L. States and Anna V. Fisher
- 63 Unsupervised Online Assessment of Visual Working Memory in 4- to 10-Year-Old Children: Array Size Influences Capacity Estimates and Task Performance**
Shannon Ross-Sheehy, Esther Reynolds and Bret Eschman
- 75 Webcams, Songs, and Vocabulary Learning: A Comparison of In-Person and Remote Data Collection as a Way of Moving Forward With Child-Language Research**
Giovanna Morini and Mackensie Blair
- 89 A Web-Based Auditory and Visual Emotion Perception Task Experiment With Children and a Comparison of Lab Data and Web Data**
Hisako W. Yamamoto, Misako Kawahara and Akihiro Tanaka
- 102 The Community-Engaged Lab: A Case-Study Introduction for Developmental Science**
Judy Liu, Scott Partington, Yeonju Suh, Zoe Finiasz, Teresa Flanagan, Deanna Kocher, Richard Kiely, Michelle Kortenaar and Tamar Kushnir
- 114 Children's (Mis)understanding of the Balance Beam (Online Edition)**
Virginie M. L. Fillion and Sylvain Sirois
- 122 Quantifying Everyday Ecologies: Principles for Manual Annotation of Many Hours of Infants' Lives**
Jennifer K. Mendoza and Caitlin M. Fausey
- 141 A Global Perspective on Testing Infants Online: Introducing ManyBabies-AtHome**
Lorijn Zaadnoordijk, Helen Buckler, Rhodri Cusack, Sho Tsuji and Christina Bergmann
- 148 "May I Grab Your Attention?": An Investigation Into Infants' Visual Preferences for Handled Objects Using Lookit as an Online Platform for Data Collection**
Christian M. Nelson and Lisa M. Oakes

- 157 ***Organizing the Methodological Toolbox: Lessons Learned From Implementing Developmental Methods Online***
Jonathan F. Kominsky, Katarina Begus, Ilona Bass, Joseph Colantonio, Julia A. Leonard, Allyson P. Mackey and Elizabeth Bonawitz
- 171 ***Natural Variability in Parent-Child Puzzle Play at Home***
Nicole Pochinki, Dakota Reis, Marianella Casasola, Lisa M. Oakes and Vanessa LoBue
- 183 ***The Negative Impact of Noise on Adolescents' Executive Function: An Online Study in the Context of Home-Learning During a Pandemic***
Brittney Chere and Natasha Kirkham
- 199 ***A Framework for Online Experimenter-Moderated Looking-Time Studies Assessing Infants' Linguistic Knowledge***
Desia Bacon, Haley Weaver and Jenny Saffran
- 216 ***Agreement and Reliability of Parental Reports and Direct Screening of Developmental Outcomes in Toddlers at Risk***
Juan Giraldo-Huertas and Graham Schafer
- 235 ***Advances in Behavioral Remote Data Collection in the Home Setting: Assessing the Mother-Infant Relationship and Infant's Adaptive Behavior via Virtual Visits***
Eunkyung Shin, Cynthia L. Smith and Brittany R. Howell
- 244 ***Designing Virtual, Moderated Studies of Early Childhood Development***
Liesbeth Gijbels, Ruofan Cai, Patrick M. Donnelly and Patricia K. Kuhl
- 264 ***Parent-Infant Interaction Tasks Adapted for Remote Testing: Strengths, Challenges, and Recommendations***
Shira C. Segal and Margaret C. Moulson
- 271 ***Baby's Online Live Database: An Open Platform for Developmental Science***
Masaharu Kato, Hirokazu Doi, Xianwei Meng, Taro Murakami, Sachiyo Kajikawa, Takashi Otani and Shoji Itakura
- 277 ***Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task***
Lydia Paulin Schidelko, Britta Schünemann, Hannes Rakoczy and Marina Proft
- 284 ***A Tale of Three Platforms: Investigating Preschoolers' Second-Order Inferences Using In-Person, Zoom, and Lookit Methodologies***
Elizabeth Lapidow, Tushita Tandon, Mariel Goddu and Caren M. Walker
- 294 ***Feasibility of Remote Performance Assessment Using the Free Research Executive Evaluation Test Battery in Adolescents***
Isis Angelica Segura and Sabine Pompéia
- 305 ***A Contactless Method for Measuring Full-Day, Naturalistic Motor Behavior Using Wearable Inertial Sensors***
John M. Franchak, Vanessa Scott and Chuan Luo
- 320 ***Comparing Face-to-Face and Online Data Collection Methods in Preterm and Full-Term Children: An Exploratory Study***
Paige M. Nelson, Francesca Scheiber, Haley M. Laughlin and Ö. Ece Demir-Lira
- 331 ***Remote Research Methods: Considerations for Work With Children***
Michelle M. Shields, Morgan N. McGinnis and Diana Selmeczy

- 338 *Children's Learning of Non-adjacent Dependencies Using a Web-Based Computer Game Setting***
Mireia Marimon, Andrea Hofmann, João Veríssimo, Claudia Männel, Angela D. Friederici, Barbara Höhle and Isabell Wartenburger
- 353 *Moderated Online Data-Collection for Developmental Research: Methods and Replications***
Aaron Chuey, Mika Asaba, Sophie Bridgers, Brandon Carrillo, Griffin Dietz, Teresa Garcia, Julia A. Leonard, Shari Liu, Megan Merrick, Samaher Radwan, Jessa Stegall, Natalia Velez, Brandon Woo, Yang Wu, Xi J. Zhou, Michael C. Frank and Hyowon Gweon
- 366 *A New Look at Infant Problem-Solving: Using DeepLabCut to Investigate Exploratory Problem-Solving Approaches***
Hannah Solby, Mia Radovanovic and Jessica A. Sommerville
- 384 *Creating a Corpus of Multilingual Parent-Child Speech Remotely: Lessons Learned in a Large-Scale Onscreen Picturebook Sharing Task***
Fei Ting Woon, Eshwaaree C. Yogarajah, Seraphina Fong, Nur Sakinah Mohd Salleh, Shamala Sundaray and Suzy J. Styles
- 393 *Remote Testing of the Familiar Word Effect With Non-dialectal and Dialectal German-Learning 1–2-Year-Olds***
Bettina Braun, Nathalie Czeke, Jasmin Rimpler, Claus Zinn, Jonas Probst, Bastian Goldlücke, Julia Kretschmer and Katharina Zahner-Ritter
- 411 *Tracking Infant Development With a Smartphone: A Practical Guide to the Experience Sampling Method***
Marion I. van den Heuvel, Anne Bülow, Vera E. Heininga, Elisabeth L. de Moor, Loes H. C. Janssen, Mariek Vanden Abeele and Myrthe G. B. M. Boekhorst
- 421 *Bringing Home Baby Euclid: Testing Infants' Basic Shape Discrimination Online***
Agata Bochynska and Moira R. Dillon
- 431 *Disruption Leads to Methodological and Analytic Innovation in Developmental Sciences: Recommendations for Remote Administration and Dealing With Messy Data***
Sheila Krogh-Jespersen, Leigha A. MacNeill, Erica L. Anderson, Hannah E. Stroup, Emily M. Harriott, Ewa Gut, Abigail Blum, Elveena Fareedi, Kaitlyn M. Fredian, Stephanie L. Wert, Lauren S. Wakschlag and Elizabeth S. Norton
- 443 *Implementing Remote Developmental Research: A Case Study of a Randomized Controlled Trial Language Intervention During COVID-19***
Ola Ozernov-Palchik, Halie A. Olson, Xochitl M. Arechiga, Hope Kentala, Jovita L. Solorio-Fielder, Kimberly L. Wang, Yesi Camacho Torres, Natalie D. Gardino, Jeff R. Dieffenbach and John D. E. Gabrieli
- 463 *Comparing Online Webcam- and Laboratory-Based Eye-Tracking for the Assessment of Infants' Audio-Visual Synchrony Perception***
Anna Bánki, Martina de Eccher, Lilith Falschlehner, Stefanie Hoehl and Gabriela Markova
- 482 *Remote Data Collection During a Pandemic: A New Approach for Assessing and Coding Multisensory Attention Skills in Infants and Young Children***
Bret Eschman, James Torrence Todd, Amin Sarafraz, Elizabeth V. Edgar, Victoria Petrulla, Myriah McNew, William Gomez and Lorraine E. Bahrick
- 496 *Zoom, Zoom, Baby! Assessing Mother-Infant Interaction During the Still Face Paradigm and Infant Language Development via a Virtual Visit Procedure***
Nancy L. McElwain, Yannan Hu, Xiaomei Li, Meghan C. Fisher, Jenny C. Baldwin and Jordan M. Bodway



Editorial: Empirical Research at a Distance: New Methods for Developmental Science

Sho Tsuji^{1,2*}, Dima Amso³, Rhodri Cusack⁴, Natasha Kirkham⁵ and Lisa M. Oakes⁶

¹ International Research Center for Neurointelligence, The University of Tokyo Institutes for Advanced Study, The University of Tokyo, Tokyo, Japan, ² Institute for AI and Beyond, The University of Tokyo, Tokyo, Japan, ³ Developmental Cognitive Neuroscience Laboratory, Department of Psychology, Columbia University, New York, NY, United States, ⁴ Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland, ⁵ Centre for Brain and Cognitive Development, Department of Psychological Sciences, Birkbeck, University of London, London, United Kingdom, ⁶ Center for Mind and Brain, Department of Psychology, University of California, Davis, Davis, CA, United States

Keywords: online testing, developmental psychology, remote testing, new methods, child development, COVID-19

Editorial on the Research Topic

Empirical Research at a Distance: New Methods for Developmental Science

INTRODUCTION

The COVID-19 pandemic presented many challenges for the research community. The collection of papers in this Research Topic illustrate how developmental scientists met those challenges and created clever and innovative methods to continue research when it was not safe to have children and families physically in the lab. Soon after labs were closed by universities and institutions, developmental scientists were scheduling video conferences with children to collect data, programming web-based procedures for participation, and considering ways to reevaluate previously collected data. The papers presented here demonstrate how the community continued to conduct research even though we were not able to work directly with our participants.

These papers reflect a diverse set of approaches to studying a wide range of content. They not only demonstrate the effectiveness (or ineffectiveness) of these methods, but also engage discussion on their drawbacks and gains. Are there advantages of new online paradigms with respect to increasing our reach to wider participant pools than usually recruited? If so, do these advantages outweigh the very real disadvantages of a decrease in the precision of measurements (e.g., not being able to control for distraction in the testing environment)? What criteria would our field need to develop for the adoption of such new methods (e.g., privacy concerns, ethical considerations)? Liu et al. discuss the benefits of reaching out into the community to find collaboration and to engage with participants regarding research ethics and values.

This Editorial is organized as follows. First, we describe the wide range of methods and measures adopted, illustrating how the move to collecting data at a distance did not restrict the ways we conducted research or the questions we asked. Next, we describe efforts to directly compare the results of data collected online (both supervised and unsupervised) to data collected in person. This Research Topic of papers reveals both findings that are context-independent (i.e., the same pattern is observed regardless of how the data were collected) and context-dependent (i.e., different patterns are observed in online vs. in-person data). In addition, these papers address questions of how procedures need to be modified, differences in data quality, and what measures can and cannot be assessed in different data collection contexts. We then present “lessons learned” and advice for best practices. We suspect that developmental scientists will continue to collect data at a distance, and the work presented here can provide guidelines to ensure that future efforts produce high quality work. Finally, we discuss what online remote research can offer—and what it cannot—as the field moves forward.

OPEN ACCESS

Edited and reviewed by:

Katharina J. Rohlfing,
University of Paderborn, Germany

*Correspondence:

Sho Tsuji
shotsuji@ircn.jp

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 08 May 2022

Accepted: 12 May 2022

Published: 25 May 2022

Citation:

Tsuji S, Amso D, Cusack R, Kirkham N
and Oakes LM (2022) Editorial:
Empirical Research at a Distance:
New Methods for Developmental
Science. *Front. Psychol.* 13:938995.
doi: 10.3389/fpsyg.2022.938995

THE RANGE OF METHODS AND MEASURES

Researchers who were unable to collect data in person adopted a number of different approaches to continue their research. Some explored ways of analyzing previously collected data. For example, Solby et al. applied neural network analyses to archival data on infants' problem solving abilities. Mendoza and Fausey provide guidance for manually annotating children's everyday experiences from data in repositories. Many others began developing or using tools for collecting data remotely. For some researchers this meant creating versions of their experimental procedures that could be administered in a moderated video conference (e.g., using a platform such as Zoom). Other researchers used or developed procedures for unsupervised data collection, in which the participants or families used their own computers or equipment provided by the researchers to collect data in their own homes (e.g., the online experimental platform Gorilla). We next describe the work conducted with moderated and unmoderated procedures.

Moderated Procedures

Many of the papers in this collection provide examples of moderated or synchronous remote data collection. In these procedures, participants typically make an appointment and meet with the researcher remotely *via* a video conferencing platform. This is essential when the experimental paradigm requires that children interact with and respond to instructions given by a researcher. Researchers used this approach to investigate a wide range of questions, including school aged children's solutions to balance beam problems (Filion and Sirois), young children's performance on traditional false belief tasks (Schidelko et al.), mother-infant interaction (McElwain et al.), and standardized cognitive functioning assessments like Mullen or Bayleys (Krogh-Jespersen et al.).

Moderated sessions also can be less structured in order to capture more "naturalistic" behaviors at home. Moderated sessions have been used to record free-play with parents and infants (Shin et al.; Segal and Moulson), puzzle play with preschoolers and parents (Pochinki et al.), and eating behaviors at mealtime (Venkatesh and DeJesus). In a semi-structured approach, Woon et al. recorded parents reading a book with their infants or toddlers, using the screen sharing feature on Zoom to present the same book to all participants.

There are also examples of researchers conducting multi-session and training studies using fully remote experimenter-moderated sessions. Bambha and Casasola had an experimenter meet with children on Zoom, every week for 5 weeks, to deliver a spatio-cognitive and visuo-motor skill training protocol. Ozernov-Palchik et al. delivered a fully remote language intervention and assessed its impact. Both papers discuss the challenges and strengths of such a multi-session remote approach.

Because they allow for better monitoring of caregiver and child variables, some researchers chose moderated sessions for tasks that could have been conducted in unmoderated sessions, including looking-time procedures with young children

(Bacon et al.; Chuey et al.; Morini and Blair) and monitoring children completing tasks using Qualtrics (Qualtrics, 2022) and other software (Segura and Pompéia; Vales et al.). Researchers choose moderated sessions for a variety of reasons including ease of setting up the procedure, a desire for more control, targeting a particular population, and comparing results between moderated and unmoderated studies.

Unmoderated Procedures

Many researchers elected to conduct unmoderated or asynchronous remote data collection, especially for screen-based, non-interactive experimental tasks. Platforms such as Lookit (Scott and Schulz, 2017) facilitated the administration of infant looking time tasks, in which researchers can set up stimuli to present to infants or young children and record their looking to those stimuli. Platforms such as Gorilla (Anwyl-Irvine et al., 2020) or LabVanced (Finger et al., 2017) allow researchers to design and program experiments to collect reaction time and accuracy as children press keys on their computer keyboard in response to stimuli presented on the monitor. These unmoderated procedures have the advantage that participants can log into an experimental program over a web browser and participate in an experiment in their own time by following the screen prompts. Oftentimes, the experimental software allows tight control over experimental variables like stimulus presentation and timing. They have the disadvantage that there is no experimenter to direct the parent or child, to make sure that the setting and recording is optimal, and to ensure compliance with the task. Nevertheless, several papers in this Research Topic demonstrate that these can be effective procedures.

For example, Nelson and Oakes demonstrated that infants' visual preference can be examined using the unmoderated platform Lookit and labor-intensive off-line coding. Others presented procedures that code looking automatically, either online or after data recording. Using the built-in webcam-based automatic eye tracking feature of LabVanced, Bánki et al. conducted an online eye tracking study to assess 4- to 6-month-old infants' sensitivity to audio-visual synchrony. Braun et al. developed an app for the iPad that recorded videos of toddlers' responses to images corresponding to familiar and unfamiliar words. Children's looking time was later analyzed using a combination of human coding and neural networks. Eschman et al. described how existing deep learning tools for face recognition can be adapted to automatically code eye gaze from recorded sessions.

Other kinds of responses can be recorded in unmoderated sessions. Marimon et al. used LabVanced to collect reaction time from 3½ to 8-year-old children who responded with button presses to assess their sensitivity to non-adjacent dependencies in linguistic stimuli. Ross-Sheehy et al. used Gorilla to record button presses from 4- to 10-year-old children in a change detection task as a measure of their visual working memory. Chere and Kirkham investigated executive functions in contexts of noise with 11–18 year-olds on Gorilla, in which both accuracy and reaction time measures were collected.

Another approach to unmoderated research was to train caregivers to collect data in their homes or during their daily lives.

Franchak et al. demonstrated how they could study infants' body positions using a set of wearable inertial sensors delivered to the infants' homes and applied by the parents and developing neural-network based analyses of these body postures. Van den Heuvel et al. discussed the value and pitfalls of experience sampling methods (ESM) using smartphones to gather data on infants and their families.

COMPARISON OF IN-PERSON VS. REMOTE DATA COLLECTION

Regardless of the particular data collection procedure, an important question is how the results of data collected remotely compares to data collected in-person. Given the lack of control in the testing environment—and the presence of many more distractions than in the lab—it is not immediately obvious whether data collected remotely will yield the same results as data collected in person in a lab setting. This central question was explored by a large proportion of authors, and the results were mixed.

One issue is simply whether the quality of the data are comparable to those collected in the lab. It would not be surprising if data collected online were noisier, as there are many variables that are difficult to control (e.g., distractions, lighting, and quality of recording device). On the other hand, children may be more comfortable at home, and thus online data collection may actually have less noise.

For some procedures, the data quality for online studies was quite good. Bacon et al. reported that data loss in a looking-while-listening task was similar to that observed in the lab. Morini and Blair reported similar numbers of trials from preschoolers in the lab and tested online in a vocabulary learning task (using looking as a measure) with preschoolers. But others reported poor data quality from online sessions, for example when remotely conducting eye-tracking (Bánki et al.) or recording audio responses (Gijbels et al.). There are remedies to some sources of poor data quality, however. Gijbels et al. for example, provided children with wearable audio recorders (LENA, Xu et al., 2009) to obtain higher quality audio data than can be obtained from Zoom recordings.

A second issue is whether the same patterns of results are observed in both contexts. Several studies found no differences between data collected in remote and in-lab sessions. Attempts to replicate previously collected (and often published) findings from lab-based research were successful. For example, in a moderated task using Gorilla, Yamamoto et al. replicated previous findings from lab-based studies for children's emotion perception in auditory and visual stimuli. Vales et al. used a Qualtrics task in a moderated session and replicated previously reported findings about 4- to 6-year-old children's semantic knowledge. Schidelko et al. reported results from online false belief tasks with preschoolers that replicated previous findings. However, Bochynska and Dillon conducted a visual preference study with infants using Lookit, and did not replicate findings on infant shape discrimination from data collected in the lab.

Others directly compared data collected in the lab and online. In some cases the procedures and methods were very similar, as

were the results. For example, Segura and Pompéia compared results when 9- to 15-year-old children were administered a battery of executive function tasks by an experimenter, either in person or moderated online, and observed no differences in performance. Morini and Blair found no differences in a looking task assessing vocabulary learning in toddlers when conducted online or in person. Silver et al. found that 2- to 3-year-old children responded similarly in a number tasks given online or in the lab. Chuey et al. replicated a number of studies of social cognition in young children using in-lab and remote testing methods. In other cases, the results differed in the two contexts. Not only did Bánki et al. find different quality of eye tracking recorded online and in person, they also obtained different patterns of results. In a comparison of performance on a second order inference task conducted in the lab, in a supervised online task, and in an unsupervised online context, Lapidow et al. observed that the online findings were weaker, and only oldest children tested show above chance performance in that context. In Bacon et al.'s looking-while-listening task (coded frame by frame through Zoom), both accuracy and reaction times showed differences from in-lab studies, with toddlers faster and more accurate in the Zoom study.

In summary, although some findings are robust to differences in testing context, others are not. This observation has implications both for how we think about specific findings—and whether or not they are robust and replicable—and also for what kinds of questions must be asked in a lab context and what kinds of questions can be asked utilizing remote methods.

ONLINE DATA COLLECTION CHALLENGES AND BEST PRACTICES

A significant contribution of the papers in this Research Topic is what was learned and how online remote testing can be effective, which we discuss in the following.

Adapting Procedures for Online Testing

As many of us discovered early in the COVID-19 pandemic, setting up an online study is not necessarily easy or fast. Many online platforms, such as Lookit or Labvanced, require learning new paradigm construction tools. When using less technically demanding platforms, such as Zoom, researchers discovered the importance of testing internet speed (e.g., Bacon et al.; Eschman et al.) or the limitations of some aspects of the recording for obtaining high quality data (Gijbels et al.). The challenges are not just technical, however. Researchers must consider how their tasks and procedures must be adapted for administration remotely and online. For example, Krogh-Jespersen et al. described how they adapted the Mullens, which is a standardized tool that requires using specific materials. They used parents as test administrators and adapted materials for presentation using PowerPoint, eliminating items that could not be tested remotely.

Several of the papers in this Research Topic provide guidance for the decisions researchers need to make when considering moving their task or procedures online. Kominsky et al. provide guidance to decide on whether moderated or unmoderated procedures are best, for example considering the

importance of experimenter involvement vs. the convenience of participants completing the study on their own schedule. Shields et al. provide an overview of some of the platforms available for online, unmoderated testing, which vary in their expense, the responses that can be recorded, and the ease with implementing new procedures. Braun et al. show the advantage of developing custom-build solutions, if one's research team has the technical skills.

A significant consideration is stimulus presentation. Presenting stimuli remotely is more complicated than in the lab. Some researchers send stimuli or materials home to families, and record children interacting with those materials during moderated sessions (Kominsky et al.; Silver et al.). It is more common for researchers to present stimuli over the internet during moderated or unmoderated sessions, using screen sharing, downloading stimuli onto participants' computers, or streaming on the web. The different methods have different pros and cons, including lags and dropped frames, slow internet speeds, and temporal differences. Kominsky et al. describe how researchers must balance the need for control over stimulus presentation and the quality of the stimulus presentation.

In addition, interacting with subjects online is different from in person. Experimental tasks may therefore need to be changed. Because children's attention may be more difficult to maintain during online sessions than in the lab, the recommendation is that tasks are kept short and that experimenters elicit regular responses from children (in moderated tasks) to monitor children's attention (Chuey et al.; Shields et al.).

Security Considerations

Online data collection requires that researchers consider data security. Information technology policies on University campuses frequently change, and requirements for how data collected from individuals can be stored and transmitted varies from institution to institution and from country to country. Basic questions such as what data can be collected, who has access to it, and how it is stored can be a challenge. The US has different standards and concerns than Europe, which may make collecting data in both environments difficult (Zaadnoordijk et al.).

As a result, researchers must consider carefully the platforms they adopt to collect data with children and families. Chuey et al. provide pros and cons of several popular video conferencing platforms for the purposes of data collection. For example, Zoom has security features such as real-time encryption, the ability to require a passcode and enable waiting rooms (Gijbels et al.; McElwain et al.; Shin et al.). It can allow researchers to record sessions directly onto their local harddrives (Bacon et al.; Segal and Moulson; Venkatesh and DeJesus), or to have participants record their sessions on their own hard drives (to avoid lags; Morini and Blair). In this second case, the researcher has to have a way to securely transfer the recording from the participant's computer to the researcher's computer. Regardless of how the research team solves these issues, online testing raises privacy issues as it often involves creating recordings that show parts of the participants' homes. That is, although online data collection can provide insight into children's environment (Chere and Kirkham) and how children behave while at home

(Pochinki et al.), it also exposes the researcher to a new level of privacy and security concerns.

Involving Caregivers in the Study Process

When testing participants online, the opportunities for instructions are more limited than in the lab, even in moderated sessions. In the absence of an experimenter and a lab setting, parents and other caregivers often play an important role in order to ensure adequate study setting and control. Shields et al. provide suggestions for how to involve parents in this way, and researchers in this Research Topic often emphasize the role of parents as active co-researchers (e.g., Eschman et al.; Zaadnoordijk et al.). How this could best be achieved depends on the required caregiver contributions and the task format; for instance, Krogh-Jespersen et al. emphasize the importance of creating rapport between caregivers and researchers in longer, moderated tasks, while shorter, unmoderated experimental protocols might especially benefit from clear instructions (Shin et al.).

For the latter, checklists and tutorial videos are recommended to ensure parents set up their home study environment in a way that minimizes interruption and distraction (Shin et al.). Another technique that researchers put forward is to do pre-study sessions with parents, including technical and equipment tests to check that parents use the correct devices and that quality of stimuli and internet speed were sufficient (Eschman et al.; Morini and Blair). What each of these examples illustrate is that involving and training the parent can have a positive impact on data quality and the overall success of remote data collection.

THE PROMISES OF REMOTE TESTING

The promise of remote data collection is enticing. Developmental scientists have long struggled with collecting ample sample sizes, as well as having samples that are diverse and representative of all children. In addition, remote data collection is more accessible to researchers who have limited space and resources to collect in-person data. Thus, although the COVID-19 pandemic motivated many to collect data online out of necessity, it is likely that many researchers will continue to collect data remotely even after it is possible to collect data in person.

The shift to online testing made it possible for developmental scientists to ask and answer questions that are difficult or impossible to address in-person in a lab. Remote research provides insights into children's lives at home that is only possible with remote testing. Pochinki et al., for example, showed how with remote testing we gain understanding into the kinds of puzzles preschoolers play with their parents, and the kinds of behaviors mothers and preschoolers engage in during that play. Chere and Kirkham assessed the impact of noisy home environments on executive functioning in adolescents, illustrating how remote testing can tap into aspects that are hard to assess in the lab. Franchak et al. collected extensive data about motor behavior during naturalistic interactions at home by sending home equipment and instructing parents how to use it at home. These papers illustrate how remote testing gives

us insight into development in context in a way that lab-based research cannot.

Online methods also have promise for developmental screening, which is expensive for health services to conduct in-person. Giraldo-Huertas and Schafer compared a standardized developmental screening with a parental measure that could in principle be administered online. Nelson et al. directly compared how pre- and full-term children performed on standardized and experimental cognitive assessments at 4 and 5 years of age in person and online. They found no differences as a function of format on 5 of 8 tasks and found that there were no effects of format for children at risk.

One still at least partly unfulfilled promise of online data collection is a more global reach and inclusivity. For instance, Lookit, the main platform for infant looking time studies, is primarily available for families living in English language environments and under US data protection laws. Nevertheless, we think that this problem is more surmountable in online than in-lab settings, and indeed, projects like ManyBabies-AtHome (Zaadnoordijk et al.) aim to globally broaden access to relevant software and data management options. A related problem is recruitment, where again research recruiting English-speaking and US-based families can profit from quickly evolving platforms such as ChildrenHelpingScience, with equivalents for other areas only sparsely available (but see Kinder Schaffen Wissen for German speakers). Kato et al. tackle the problem of creating a database for recruiting infants and storing data online in Japan, including the creation of a researcher consortium to manage such efforts. Another concern for inclusivity in online studies is the necessity of a stable internet connection and a device to participate in studies. A lot of work still needs to be done to overcome these problems, but this Research Topic assembles some suggestions for solutions, such as lending participants a Wifi tool or hotspot or refer them to public places that offer free internet, or to create tasks that allow participants to participate over their mobile phone as opposed to a webcam-enabled computer (Shin et al.). Thus, while remote data collection is still not as global and inclusive as we might have imagined at the outset of the pandemic, the research community suggests and has

started implementing concrete and attainable solutions toward this goal.

Even if researchers will solve the practical problems of testing a diverse subject population, online testing does not guarantee that diverse samples will be automatically recruited. For example Bacon et al. deliberately tested the idea that they could recruit a more diverse sample online by using microtargeting Facebook ads. However, this study also illustrates that although in principle online testing provides access to populations who would not ordinarily come to the lab (e.g., they live too distant), it takes effort and care to recruit more diverse populations, just as it would to recruit those samples for in-person testing. Liu et al. demonstrate the effectiveness of community engaged labs for recruiting diverse samples.

CONCLUSION

Research at a distance is here to stay for developmental science. The collection of papers in this Research Topic illustrate many of the ways that methods and procedures can be adapted for remote administration. The papers provide models for solutions to common problems, and will help researchers in the future make decisions about how to conduct empirical research at a distance to answer key questions in developmental science.

AUTHOR CONTRIBUTIONS

ST and LO wrote the manuscript. All authors made substantial, direct, and intellectual contributions to this work and approved it for publication.

FUNDING

This work was supported by an ERC Advanced Grant ERC-2017-ADG, FOUNDCOG, 787981 to RC, as well as a JSPS Grant-in-aid for Specially Promoted Research (20H05617), JSPS Grant-in-aid for Transformative Research Areas (20H05919), and a JST-ActX grant in the research area AI powered Research Innovation/Creation awarded to ST.

REFERENCES

- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: an online behavioral experiment builder. *Behav. Res. Method.* 52, 388–407. doi: 10.3758/s13428-019-01237-x
- Finger, H., Goeke, C., Diekamp, D., Standvoß, K., and König, P. (2017). “LabVanced: a unified JavaScript framework for online studies,” in *International Conference on Computational Social Science*. Cologne.
- Qualtrics (2022). *Qualtrics and All Other Qualtrics Product or Service Names Are Registered Trademarks or Trademarks of Qualtrics*. Provo, UT. Available online at: <http://www.qualtrics.com> (accessed May 8, 2022).
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Xu, D., Yapanel, U., and Gray, S. (2009). *Reliability of the LENA Language Environment Analysis System in Young Children's Natural Home Environment*. Boulder, CO: LENA Foundation.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Tsuji, Amso, Cusack, Kirkham and Oakes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



From Lab to Zoom: Adapting Training Study Methodologies to Remote Conditions

Valerie P. Bambha* and Marianella Casasola

Play and Learning Lab, Department of Psychology, Cornell University, Ithaca, NY, United States

OPEN ACCESS

Edited by:

Dima Amso,
Brown University, United States

Reviewed by:

Sarah Elizabeth Rose,
Staffordshire University,
United Kingdom
John Franchak,
University of California,
Riverside, United States

*Correspondence:

Valerie P. Bambha
vpb27@cornell.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 13 April 2021

Accepted: 29 June 2021

Published: 19 July 2021

Citation:

Bambha VP and Casasola M (2021)
From Lab to Zoom: Adapting Training
Study Methodologies to Remote
Conditions.
Front. Psychol. 12:694728.
doi: 10.3389/fpsyg.2021.694728

Training studies extend developmental research beyond single-session lab tasks by evaluating how particular experiences influence developmental changes over time. This methodology is highly interactive and typically requires experimenters to have easy, in-person access to large groups of children. When constraints were placed on in-person data collection due to the COVID-19 pandemic, administering this study format in the conventional manner became unfeasible. To implement this type of research under these new circumstances, we devised an alternative approach that enabled us to conduct a live, multi-session training study using a diverse array of activities through an online interface, a task necessitating creative problem solving, since most existing remote methodologies either rely on unsupervised methods or have been limited to single sessions and restricted to a limited number of tasks. The current paper describes the technological and practical adaptations implemented in our online training study of 118 4- and 5-year-old children from a geographically diverse sample. An experimenter interacted with the children once a week for 5 weeks over Zoom. The first and final sessions were dedicated to collecting baseline and post-test measures, while the intermediate 3 weeks were structured as a training designed to teach children specific spatial-cognitive and visuo-motor integration skills. The assessments and training contained image-filled spatial tasks that experimenters shared on their screen, a series of hands-on activities that children completed on their own device and on paper while following experimenters' on-screen demonstrations, and tasks requiring verbal indicators from the parent about their child's response. The remote nature of the study presented a unique set of benefits and limitations that has the potential to inform future virtual child research, as our study used remote behavioral methods to test spatial and visuo-motor integration skills that have typically only been assessed in lab settings. Results are discussed in relation to in-lab studies to establish the viability of testing these skills virtually. As our design entailed continual management of communication issues among researchers, parents, and child participants, strategies for streamlined researcher training, diverse online recruitment, and stimuli creation are also discussed.

Keywords: Zoom, preschool, training study, remote testing, learning, spatial skills, visuo-motor integration

INTRODUCTION

The integral role of experience is studied in lab settings through *training studies*, a multi-session methodology in which experimenters first assess children at baseline on measures of interest and then, in subsequent lab sessions manipulate the types of input and experiences the children receive based on their assigned condition. At the end of the study, children are tested again on the same measures as at baseline and the results are analyzed to determine differential patterns of success across the training conditions. Training studies have been effectively applied to a wide variety of developmental domains, such as social cognition, language development, mathematical cognition, spatial skills, working memory, and the development of positive psychological traits such as optimism (Hale and Tager-Flusberg, 2003; Uttal et al., 2013; Hofmann et al., 2016; Gade et al., 2017; Malouff and Schutte, 2017; Casasola et al., 2020; Mix et al., 2020). Their study design makes it possible to draw causal conclusions about what specific aspects of these controlled experiences lead to improvements in particular kinds of skills.

Training studies typically require that experimenters have access to large groups of children in person over an extended period of time. However, even in ordinary circumstances this type of recruitment is challenging, as researchers must pull and retain a sufficient sample from the limited geographic areas surrounding their universities. Following the beginning of the March 2020 social distancing restrictions in the United States due to the COVID-19 pandemic, in-person data collection at most universities was stopped completely or severely restricted. As of May 2021, in-person data collection remains limited at most institutions across the country, especially with vulnerable populations such as children. It therefore became necessary for child researchers to think of creative solutions to translate their methods to online platforms. This problem was particularly messy for experimenters who wanted to maintain the hands-on and longitudinal nature of training studies.

While remote methodologies for developmental research exist that predate the COVID-19 era, they are not the most suitable for training studies that seek to provide children with controlled, multimodal, and interactive experiences that target a range of skills over multiple sessions. Most established remote methodologies were intended to be implemented in the absence of a live experimenter (Scott and Schulz, 2017; Rhodes et al., 2020). For example, parents and children using the online Lookit platform¹ from Scott and Schulz (2017) never interact directly with an experimenter. Instead, infants and children watch videos of prerecorded stimuli, and their eye gaze is captured and saved through their webcam. Likewise, studies implemented through Discoveries Online,² an unmoderated interface designed for verbal children ages three and older, participants make selections on their screen based on study narratives and animations (Rhodes et al., 2020). Although these methods allow families to complete sessions at their convenience,

they do not lend themselves well to tasks requiring the child's active participation, as during Lookit tasks the child simply watches the screen and is not able to interact directly with anything they see, and children participating in studies from Discoveries Online are constrained to actions that can be elicited with little setup and explanation, such as pressing a button on the screen or discussing a story with their parent in a naturalistic setting (Scott and Schulz, 2017; Rhodes et al., 2020). Because there is no live experimenter in either methodology, there is no mechanism in place to ensure that the instructions are followed, the child remains engaged and fixated on the screen, the camera angles stay in focus, and the data upload correctly. This last point is particularly pertinent because without an experimenter present to assume the responsibility of recording and administering the session, approximately 35% of the Lookit videos analyzed in Scott and Schulz (2017) were unusable, the majority due to missing or incomplete video data. Additionally, although Rhodes et al. (2020) reported a low level of parental interference in the studies conducted on their Discoveries Online platform, it is important to note that their studies that were not explicitly about parent-child interactions and were intentionally designed to require as little parental involvement as possible. Though such a setup reduces the risk of parental interference, it is not well suited for the goals of a training study that require child engagement in specific activities over several sessions.

Moreover, one of the only existing empirical studies that has used live videoconferencing to interface with children, Roseberry et al. (2014), a language learning study that examined whether social contingency would aid toddlers' ability to learn words from digital applications such as Skype, was conducted in a single session in a lab setting that only used videoconferencing for a small portion of the session, and only for children in one of the three study conditions. This video chat was supplemented by a warm-up period during which the child was able to play with toys and meet the experimenters face to face, and in-person data collection methods such as eye-tracking using a physical eye-tracker. The videoconferencing component itself was also not entirely interactive, as children participated in short verbal exchanges with the experimenter at the beginning of the chat, but transitioned to passively watching and listening to the experimenter during the actual word-learning tasks (Roseberry et al., 2014). The children did not complete any participatory activities related to the word-learning task or engage the experimenter in conversation about the novel words. Established remote methodologies for developmental research have therefore mostly been applied to tasks in a narrow range of domains and modalities that are meant to capture either implicit measures or the impact of limited forms of interaction that are not directly related to the skill the child is learning.

By contrast, hands-on training studies that teach children specific skills through distinct, multimodal activities over multiple sessions have not yet been attempted in remote settings. To successfully carry out such a study, researchers would have to create study stimuli and activities that allow children to actively participate in a virtual environment, find enough families that are willing to commit to several live, online study appointments,

¹www.lookit.mit.edu

²www.discoveriesonline.org

and maintain efficient and effective communication between families and the research team, all while fostering an interactive and engaging atmosphere during the sessions themselves. The present paper details the novel approach that our research team adopted to address these obstacles in a spatial training study with 118 preschool children. In addition to being the first instance of an entirely remote training study, our study was the first of its kind to test spatial-cognitive and visuomotor integration skills, which generally rely heavily on physical materials or detailed eye-tracking methods, through behavioral methods administered through virtual interactions with experimenters. It featured baseline and post-test assessments on a variety of spatial and visuo-motor integration skills, as well as trainings with hands-on drawing activities. We will review our study's strategy for (1) participant recruitment, (2) stimuli creation and piloting, (3) study procedure and task structure, (4) long-distance research team training, and (5) parent communication. Although this approach was devised out of necessity, we believe its takeaways can be applied to future developmental studies to overcome some of the traditional recruitment limitations in the field, such as lack of geographical diversity, and expand the reach of our science. However, it is also important to note that even though our team was successful in applying remote methods and obtaining quality data, we cannot assume that the research experiences children received over Zoom is comparable to the usual in-person experience. Future work is needed to more directly compare the patterns and quality of data obtained in remote and in-person developmental studies.

PARTICIPANT RECRUITMENT

Our final sample size was 118 participants (65 girls, $M = 5.05$, $SD = 0.517$, range = 3.78–5.94 years). The study was conducted in two five-week rounds with different participants. Just over half of this sample participated in the five-week study in the summer ($N = 67$, 36 girls, $M = 5.0$, $SD = 0.510$, range = 3.78–5.90 years) while the remaining participants completed the five-week study in the fall ($N = 51$, 29 girls, $M = 5.10$, $SD = 0.51$, range = 4.16–5.94 years). Most participants ($N = 105$) were recruited from public and private Facebook groups designed for parents looking for virtual activities during the pandemic or online homeschooling resources for their children. All Facebook recruitment was handled by the first author. When asking permission to join a private Facebook group we made our intentions to advertise the study clear in our request form. Our ad contained an image and text describing the purpose, age requirements, format, length, and compensation for the study.

Interested parents replied to the lab email address, commented on the post, or messaged the first author directly. If a parent left a comment indicating that they were interested in having their child participate but did not email the lab or send a private Facebook message, the first author began communication by sending the parent a message first. After the initial contact the first author sent a follow-up email or message with more detailed information about the study, including the materials needed (two separate electronic screens were required, with a preference that one be a tablet), an explanation of the links they would be receiving

from their experimenter containing the activities, a reminder of the study format, length, and compensation, and the projected start date of the study with a request that parents send three ranked day and time preferences (in Eastern Time) for their sessions. For organizational purposes, we asked the parents to try to pick time slots at the same time and on the same day of the week each week for each of the 5 weeks. After a timeslot had been decided, the first author then connected the family with the member of the research team who would be running their sessions. Any future communication about rescheduling was coordinated by that researcher.

Our sample was geographically diverse, with 8 participants from the New England Region, 21 from the Mid-Atlantic Region, 16 from the greater Washington Metropolitan Area where the first author is from (DC, Maryland, Delaware, West Virginia, and Virginia), 10 from the Southeastern Region (North Carolina, South Carolina, Georgia, and Florida), 8 from the Southwestern Region (Arizona, Texas, Oklahoma, Arkansas, and Louisiana), 28 from the Midwest, 10 from the Rocky Mountain Region, and 15 from the Pacific Region. Two participants did not report location information.

Of the children whose caregivers reported racial demographic information, 87 were Caucasian, 14 were mixed race, 1 was American Native or Alaska Native, 1 was African American, 11 were Asian, and 1 was Native Hawaiian or other Pacific Islander. Fifteen of these children were reported as Hispanic or Latino. Of the 118 families who reported maternal education, all had graduated high school and 112 had earned at least a 4-year degree.

An additional 31 children were recruited but were not included in the final sample due to failure to begin the study after setting up a timeslot ($n = 10$), failure to complete all five sessions of the study after starting ($n = 4$), parental involvement ($n = 11$), technological difficulties ($n = 1$), or fussiness ($n = 5$).

We were able to recruit a larger sample with less attrition in the summer ($n = 80$ recruited, $n = 67$ participated) compared to the fall ($n = 72$ recruited, $n = 51$ participated), possibly because families had more free time during the summer to dedicate to our study rather than during the fall when children had the added commitment of school.

Written informed consent was obtained from a parent or guardian before the first session of the study. All procedures involving human subjects were approved by the Institutional Review Board.

Families were given a total of \$25 in electronic Amazon gift cards, \$5 after their first session and \$20 after their final session. In order to receive the full \$25 families had to complete all five sessions.

STIMULI AND PROCEDURE

Baseline and Post-test Assessments

The baseline and post-test contained a total of six assessments on a variety of spatial, language, and visuo-motor integration skills. Three of these measures were administered through

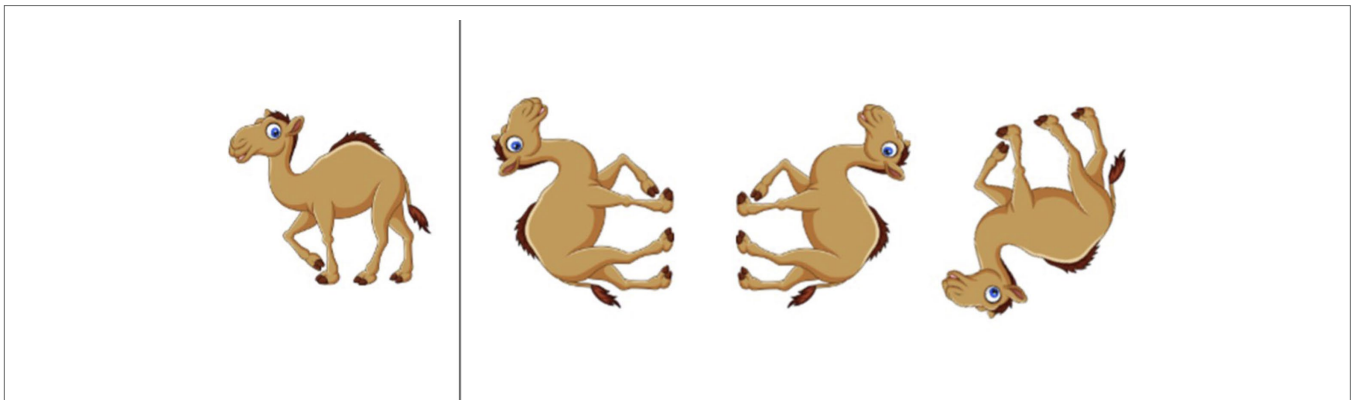


FIGURE 1 | Example trial from the mental rotation task.

Qualtrics, two were administered through another online behavioral science platform called Gorilla,³ and one was administered through virtual demonstration and a physical writing instrument and paper. We used two online platforms to more closely approximate versions of tasks that had been conducted successfully in person (Qualtrics tasks) and to administer the same version of tasks that are being used in a different ongoing online study in our lab with the comparatively older age group that was the focus of this study, setting the stage for future age comparisons (Gorilla tasks). The remaining task was adapted from a standardized visuo-motor integration task, so we used physical materials to more closely mimic its standard setup.

The first assessment was a mental rotation task based on the picture rotation task (PRT) used in Quaiser-Pohl (2003). There were two versions of the task, modeled on the two versions of the standard PRT but using different images. Children received one of the two versions at baseline and the other at post-test. During the task, they had to identify which of three rotated images exactly matched an example image. This task was administered as a Qualtrics survey that an experimenter displayed through screen share on Zoom (see **Figure 1**). Across both rounds of the study, children answered by pointing to a numbered picture and having their parent tell the experimenter which picture they chose ($n = 46$), by verbally responding themselves ($n = 71$), or by listening as the experimenter labeled each of the choices and telling them to stop when they came to the picture that they thought was the correct match ($n = 1$). The task contained three practice trials in which the child was always shown the correct answer and 12 test trials where the correct answer was not shown.

The second assessment was a novel pattern extension task that was also administered through a Qualtrics survey (see **Figure 2**). There were two versions of this task with patterns with similar structures but different specific images. Children completed one version of this task at baseline and the other at post-test. This task was conducted on a second device while the experimenter followed along by displaying the corresponding screens through screen share on Zoom. Parents were sent the

survey link prior to the testing session and had the task ready for the child to complete with the experimenter during the session. The task required children to both verbally indicate and drag and drop the three elements that came next in a series of six patterns into an answer box. They completed the task on their second device. Experimenters did not select any answers for the children during this task. Instead, children whose second device had a touchscreen ($n = 100$) used their finger to move the pictures into their correct positions in the pattern, with parental assistance as needed. Children whose second device was a laptop with no touchscreen ($n = 18$) indicated their answers by pointing to the picture on the screen and having their parent complete the drag and drop for them.

The third assessment also took place on the child's second device but was instead administered through the Gorilla platform. Children were shown a series of nine partially completed puzzles and were told to tap on the space where they thought a missing piece went (see **Figure 3**). There was no drag and drop element involved. Children whose second device had a touchscreen ($n = 100$) used their finger to tap on the matching space while those whose second device did not have a touchscreen ($n = 18$) pointed to the spot they thought was correct and their parent clicked it for them.

Children completed the fourth assessment using a physical drawing tool and paper. For this assessment, which was a modified version of the Beery Developmental Test of Visuo-Motor Integration (DTVMI) from Beery (2004), the experimenter held up a series of geometric images to the screen and had the child copy them into sheets of paper that contained tables with two rows and three columns (see **Figure 4**). The child held each page up to the screen once they had filled the table. This table had been emailed to parents the night before. There was a total of 15 progressively more difficult images for the child to copy, but the experimenter stopped early if the child was unable to draw an image or expressed a desire to stop. Children completed on average 12 drawings at baseline and 13 drawings at post-test.

For the fifth assessment, children returned to Gorilla on their second device. This task was a test of visual processing and required them to pick which of two pictures at the bottom

³www.gorilla.sc

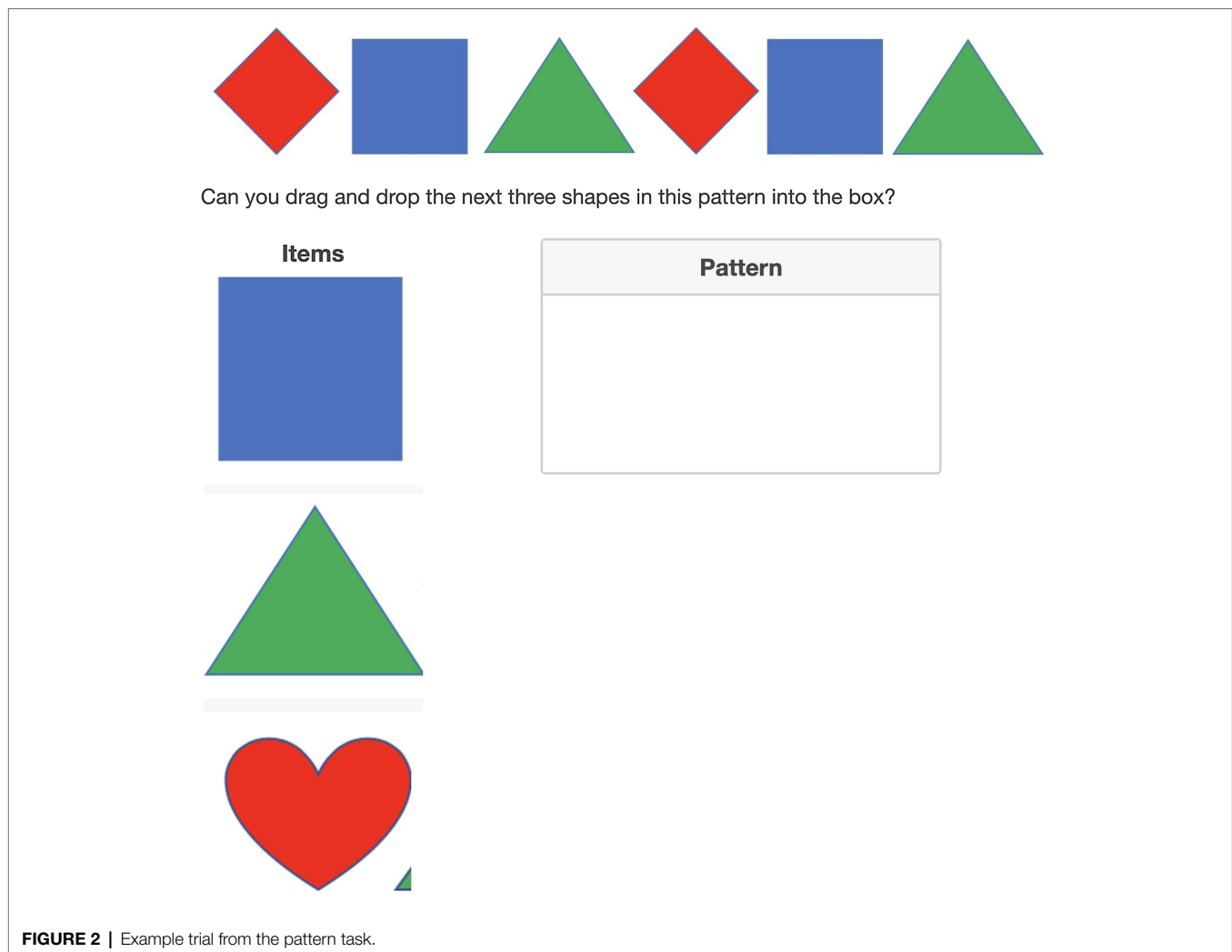


FIGURE 2 | Example trial from the pattern task.

of the screen they thought looked the most like the picture at the top of the screen (see **Figure 5**). Children whose second screen was a touchscreen ($n = 100$) used their finger to select their choice while those whose second screen was a laptop with no touchscreen ($n = 18$) pointed to their choice while their parent clicked it.

The sixth assessment was a spatial vocabulary task administered in a Qualtrics survey over screen share. There was only one version of this task and it contained items that ranged in difficulty. For the expressive part of the task, an experimenter shared pictures with various geometric shapes and spatial relations and asked the child to verbally label them (see **Figure 6**). The experimenter recorded the child's response directly into the form (24 total questions: 15 shape and 9 spatial relation). During the receptive part of the task, the experimenter provided the label and children had to select the corresponding picture (21 total questions: 9 shape and 12 spatial relation). For this part, 76 children responded by pointing to one of the numbered choices on the screen and having their parent tell the experimenter which one they pointed to

and for 29 children the experimenter verbally scanned through the numbered options and told the child to tell them to stop when they landed on the correct choice. Twelve children responded on their own.

All children were asked to complete a free draw after they completed the sixth assessment.

Depending on children's level of engagement, the experimenter sometimes presented the tasks out of order or gave the child a break to complete an extra free draw. For example, if a child started to lose focus during the interactive drag and drop task the experimenter would either move on to a less demanding task or let the child draw a picture until they regained focus and enthusiasm. There were 36 participants who were given an altered task order in this manner.

Parents were asked to take pictures of and email copies of all physical drawings their children created to the research team. Of the 118 total participants, 71 emailed all the necessary materials for both baseline and post-test. For children without emailed materials, coding was done based on the recorded Zoom video.

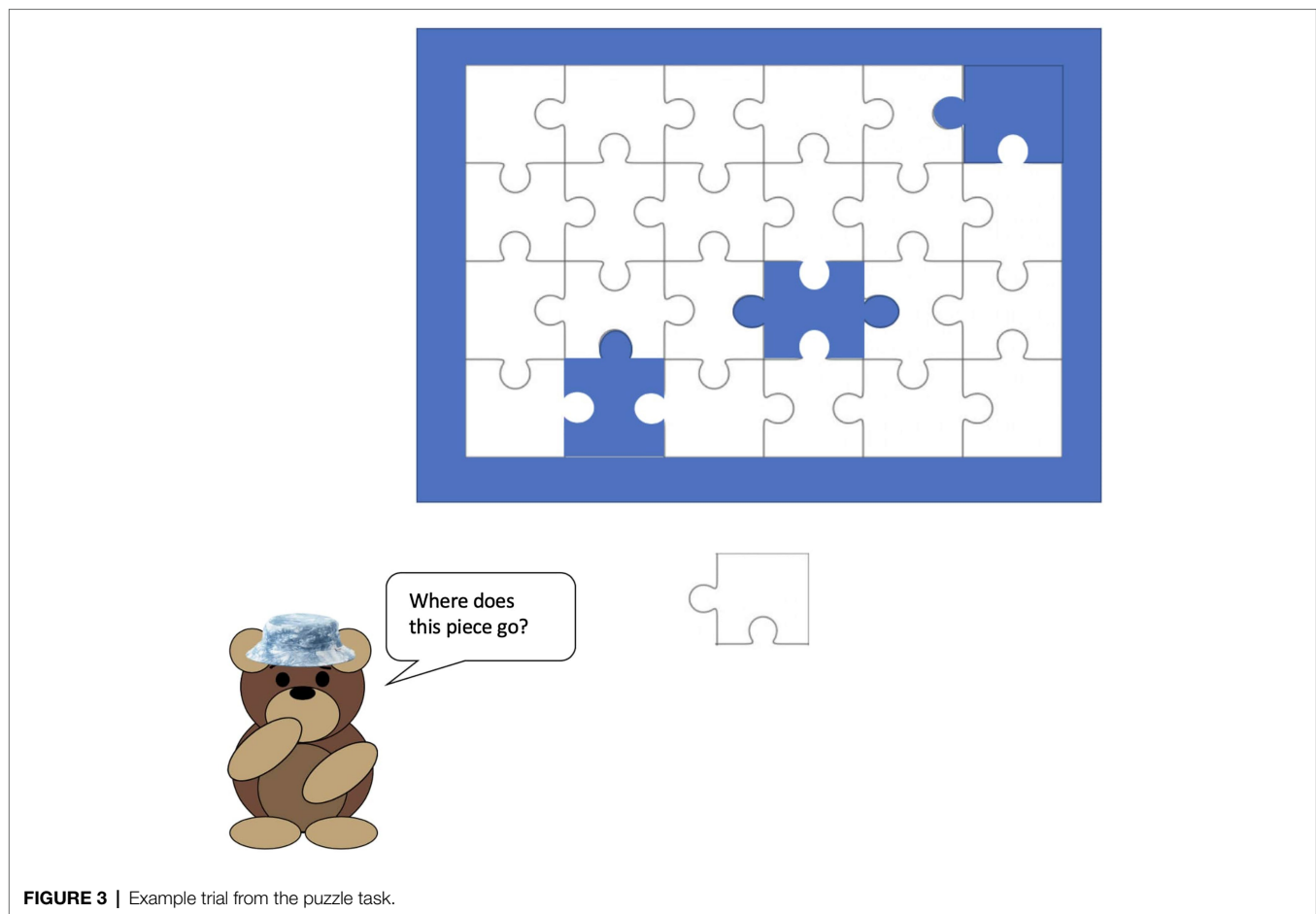


FIGURE 3 | Example trial from the puzzle task.

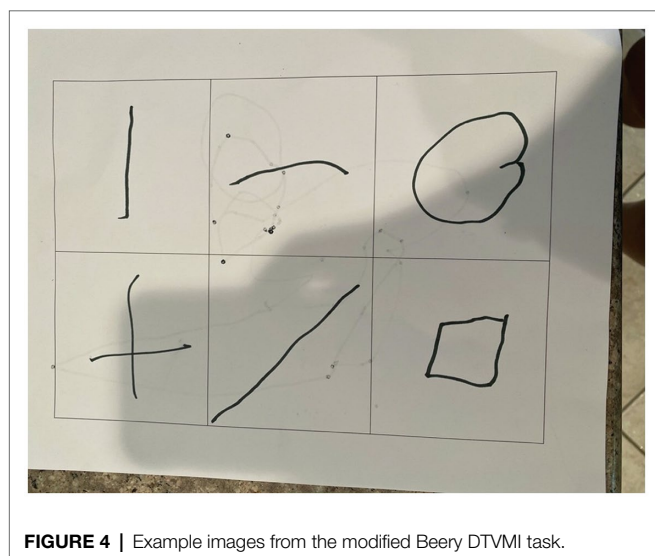


FIGURE 4 | Example images from the modified Beery DTVM task.

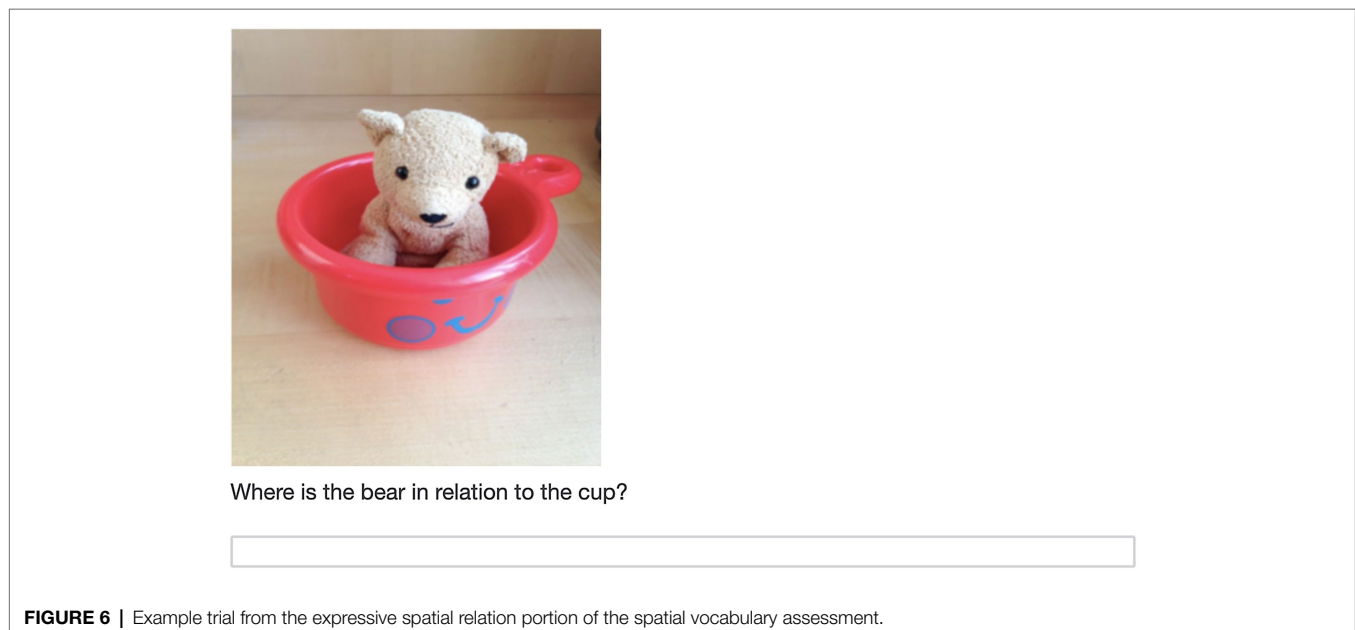
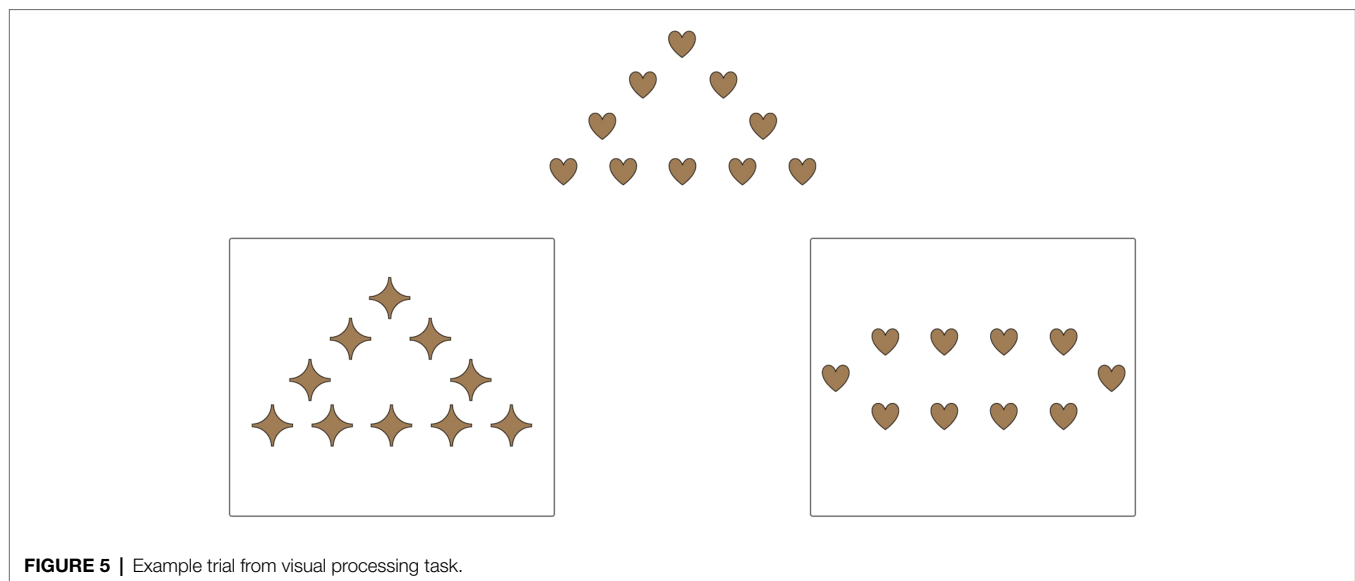
Training Activities

The training sessions were novel tasks administered through screen share by an experimenter using the Gorilla platform.

Gorilla was chosen for the demonstration because it contained a drawing tool that allowed the experimenter to draw on the screen. Children were shown and asked to trace or copy two images containing geometric shapes during the guided drawing portion of these training sessions. All children were given the same images in the same order for each session. The images for the first training session were the cat face and the penguin, the images for the second were the house with trees and the person, and the images for the third training session were the truck and the rocket (see **Figure 7**). Some children were provided with informative spatial language while they completed the art activities, and some were not.

After the two guided draws, all children completed two free draws. For the first free draw, they were instructed to draw whatever they wanted. For the second, they were told to draw whatever they wanted using as many different kinds of shapes as they could.

Parents were asked to take pictures of and email copies of all physical drawings their children created to the research team. Of the 118 total participants, 73 emailed all the necessary materials for all three trainings. For children without emailed materials, coding was done based on the recorded Zoom video.



RESEARCH ASSISTANT TRAINING

A total of 12 undergraduate research assistants, along with a hired lab manager and the graduate student principal investigator, interacted directly with the children over Zoom during data collection and assisted with behavioral coding and data processing. Three additional undergraduate research assistants worked solely on behavioral coding and data processing.

In order to maintain uniformity with such a large team that could not gather in person and that had members located in different time zones due to the unique situation created by COVID-19, we set up a series of Zoom trainings and created detailed step-by-step guides stored in our lab Box folder that

contained instructions for proper data collection protocol and links to needed materials. Research assistants also clearly marked their availability in a shared Google calendar. This organizational process proved crucial in ensuring that all the research assistants were well trained and able to carry out the protocol smoothly.

Data collection took place in two phases: summer and fall. During summer session, the first author graduate student and five undergraduate research assistants conducted sessions with 67 children, with a range of 6 to 17 participants per researcher and a total of five sessions per participant. The undergraduate students met with the graduate student over Zoom for an initial training where they were walked through the procedure over screen share and shown where the guides and materials were

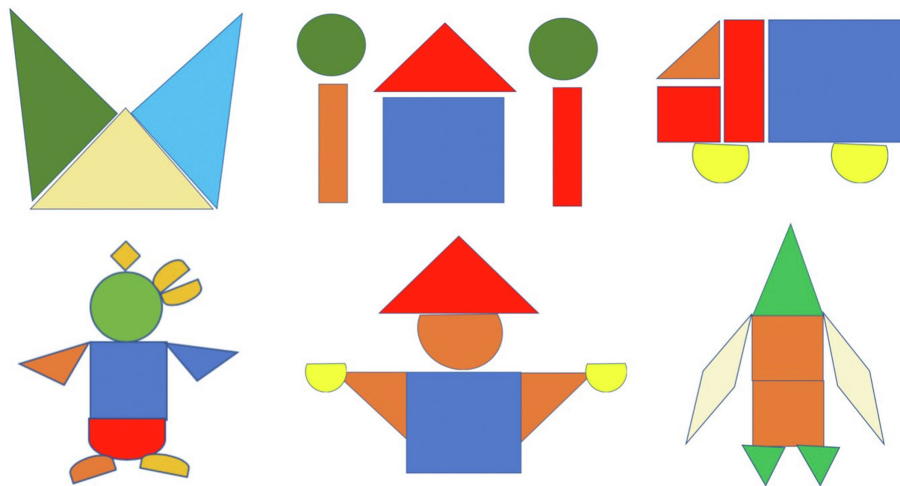


FIGURE 7 | Images completed by children during their training sessions. All children received the cat and penguin images for their first training, the house and person images for their second training, and the truck and rocket images for their third training.

stored. The first author began running participants a week before the undergraduate students, and as part of their training, the undergraduates were required to shadow the first author at least once while she ran her baseline sessions by joining as a participant on the Zoom call. As the first author remained a week ahead of the undergraduates, they were instructed to continue shadowing at least once a week in preparation for the next week's procedure and to review the first author's videoed sessions, which were also stored in the lab Box folder. The first author also remained in contact with the undergraduates *via* email, set up personal Zoom meetings to answer any questions they had about the procedure, and shadowed sessions upon request. The first author or another research assistant served as substitutes if a researcher had to miss a session for any reason.

Six new undergraduate research assistants joined the research team in the fall, and one experienced undergraduate left. Fall data collection was conducted by these undergraduate research assistants, the graduate student first author, and the lab manager for a total of 51 participants with a range of 1 to 9 participants per researcher and a total of five sessions per participant. As in the summer, the new undergraduate research assistants met with the first author over Zoom for an initial training. New research assistants were also paired with an experienced research assistant who had collected data over the summer. The purpose of these pairings was to provide new research assistants with an accessible resource who could help answer questions and troubleshoot difficulties faster than if they had to rely solely on the first author. Fall data collection began at the same time for all researchers, but the new research assistants had to shadow the experienced research assistant they were paired with at least once a week. The experienced research assistants were also expected to shadow their new research assistant at least once a week to ensure they were conducting their session correctly. The first author remained available for questions that could not be answered by the more experienced research

assistant and shadowed sessions upon request. The lab manager and the first author served as substitutes if a researcher had to miss a session for any reason.

Coding guides were created and placed in the lab Box folder detailing the necessary steps for data coding and processing for each of the tasks. Research assistants involved in data collection were expected to fulfill the remaining hours they had committed to the lab by working on coding. Three undergraduate research assistants worked only on coding and data processing. The first author set up individual Zoom meetings with each of the coders as questions arose and clarified instructions further over email.

Six undergraduate research assistants who completed an open-ended survey about their experience with remote data collection over Zoom named scheduling flexibility and geographic diversity as advantages of the approach and also said that the children seemed to be engaged with the tasks overall. The fact that the child was in their own home with their family was described as both an advantage since the child was more comfortable and did not need to warm-up as much to the experimenter, and a disadvantage since it was more difficult for the experimenter to establish authority and redirect children's attention from behind a computer screen. Respondents also said that relying on parents to redirect the child, access the needed links, and adjust camera angles was challenging, although it was easier to coordinate rescheduling sessions than when we run in-person studies in the lab. They also wrote that technical difficulties arose occasionally for both experimenters and families, but rarely significantly impacted the sessions.

COMMUNICATION WITH PARENTS

Our team found that consistent communication with parents was key for participant retention in this multi-session study.

We emailed parents the night before each of their sessions, which served both to remind them about their upcoming session and make them aware of what materials they would need, where they were located, and what they would need to do to prepare.

Communication for Baseline and Post-test Assessments

Parents were sent an email the night before their scheduled session with the Gorilla.sc link they would need to access the consent form, demographic survey, and tasks. They were also sent an additional Qualtrics link that contained the pattern task. This email also included a file with the table that the children would need for the activity requiring paper and a physical drawing tool, instructions about how to fill in their child's participant ID and other information in the forms they were sent, which device they should access each form on, and a description of the physical materials they would need for the session. Parents were also reminded that they needed to be present during the session and that they should offer encouragement, but no hints, to their children and that they would need to take pictures of anything their child physically drew and email them to the lab. Lastly, the email mentioned that they would receive a Zoom link 10–15 minutes before the scheduled start of their session, where their researcher would be watching and recording from.

Communication for Training Activities

Parents also received an email the night before each of their scheduled training sessions. Parents whose children were in the condition where they would be asked to trace images received a file with those images in this email and were instructed to copy them in pencil on two separate sheets of paper. They were told not to show these images to their children until the start of the session. All parents were told what materials they needed, were reminded to expect a Zoom link 10–15 minutes before the scheduled start of their session, and were told to take pictures of all their child's drawings and email them to the lab.

ADVANTAGES AND DISADVANTAGES OF REMOTE DATA COLLECTION FOR TRAINING STUDIES

The COVID-19 pandemic necessitated substantial adaptations to established training study methodologies. For our specific study, our main challenge was maintaining a controlled and standardized procedure across children's diverse home environments, which we could not physically manipulate, and across multiple sessions. The remote format of the study meant that the stimuli and setup of the study space, usually the responsibility of the researcher, now fell on the parent. While we tried to provide parents with detailed instructions, most parents do not have formal research background and are often

trying to set up the study in a hurry. Additionally, because families' participation is a significant service to us, we did not want to overburden them with instructions that were too cumbersome or difficult to understand. Striking the appropriate balance required trial and error, and though we were able to maintain a higher level of standardization through our interactions with parents than unsupervised remote methodologies, there were still some elements that were ultimately out of our control despite providing instructions, such as what a parent chose to say to their child during the session ($n = 2$ participants received some form of direct parental prompting about which shapes to draw during their free draw; $n = 11$ participants were excluded from analysis for at least one baseline or post-test assessment due to excessive prompting), whether the parent was present during the session at all ($n = 22$ participants had no parents present for at least one of their training sessions), technological difficulties ($n = 1$ participant was excluded from analysis of at least one task due to technological difficulties), and the child's attention ($n = 5$ participants were excluded due to fussiness). We did find that giving explicit instructions to parents about how to engage during the sessions, both in writing before the sessions and verbally during the sessions, was helpful in ensuring that the study protocol was followed. It is also worthwhile to note that although our study relied heavily on technology only one participant was excluded due to technological difficulties. Technology issues, such as slow Internet on either the experimenter's or participant's side, occasionally arose but were able to be resolved by using a Wi-Fi hotspot, restarting Zoom, or rescheduling if necessary. Internet problems were therefore not a significant impediment to data collection. However, as families were aware of the technological requirements before starting the study, it is likely that we mostly attracted families who believed they would have stable Internet.

The members of our research team were also trained extensively on what to do when they encountered issues, such as on how to verbally label answer options for the child if their parent was not present to indicate which option they had pointed to, how to provide guidance to parents about how to adjust their camera angle, suggestions for helping the parent interact with their child in a way that aligned with study protocol, solutions to common technological difficulties, and strategies for redirecting children's attention when they started to lose focus. These adaptations were necessary to maximize usable data and enabled us to offset procedural issues that arose from uncontrolled environments and that would have led to some participants being excluded from analysis of certain tasks. However, we did exclude more participants in this study compared to comparable in-person training studies carried out by our lab: $n = 31$ participants excluded in this study out of a total of 149 participants recruited compared to $n = 11$ participants excluded out of a total of 95 recruited participants in Casasola et al. (2020).

Overall, our team found that the study tasks were engaging for children and worked well virtually, though there were a few notable challenges and general observations. We observed that children were more able to independently complete

certain types of tasks than others. For example, the majority of children was able to complete the Gorilla touchscreen tasks that required a single tap on the correct answer on their own, but many seemed to struggle completing touchscreen tasks requiring more exact motor control without parental assistance. Children often became frustrated when they were not able to complete a task on their own. Children also appeared to be the most focused when they were completing a task that involved a level of participation and motor engagement from them that was neither too little nor too taxing. For example, children seemed on task when they were drawing or tracing during the training sessions and the modified Beery DTVMI assessment, selecting the matching image from the three options in the mental rotation task, or answering questions about spatial vocabulary, but at times appeared to speed through the single-tap Gorilla tasks and became frustrated by pattern extension task, which required considerable manipulation of the touchscreen. These differences suggest that children engage well both when the tasks are administered by an experimenter through screen share and when they are able to manipulate physical drawing materials but can struggle maintaining focus when asked to interact directly with a touchscreen. These observations are anecdotal and should be explored further in relation to age differences and individual differences in attention, fine motor skill, and technology exposure. While the study activities and multi-session setup worked generally well for children in the age range we used, it is an open question as to whether younger children would be able to engage in a multi-session online study with these types of interactive activities. Results would be informative about the most effective remote training methodologies for teaching spatial-cognitive skills to children across a wide age range.

Using both Qualtrics and Gorilla to conduct our screen-based tasks allowed us to make comparisons about how well the two platforms hosted our baseline and post-test assessments and our interactive training activities. As a reminder, we used Qualtrics forms for the mental rotation, pattern extension, and spatial vocabulary tasks and Gorilla for the consent form and demographics, puzzle completion task, visual processing matching task, and the tracing and drawing demonstrations during the training sessions. The biggest difference between the two platforms was the amount of touch-based interactivity that was possible to integrate into each one. While the pre-made templates in Qualtrics are restricted to a few default setups (i.e., multiple choice and free answer questions, limited touch-based drag and drop matching activities) with limited aesthetic and functional customization, Gorilla has a zone feature that facilitates the creation of more complex activities in a user-friendly manner that does not require programming knowledge. This Gorilla feature was especially helpful during the training sessions because we were able to use a zone to create a space where experimenters could use their mouse to draw the study images on the screen alongside the child. We also were able to use these zones to easily create the puzzle completion and matching tasks, which required children to touch different

parts of the screen. Gorilla also has templates for the more standard question formats that are also included in Qualtrics. However, one disadvantage of Gorilla compared to Qualtrics is that there is a small fee (\$1.20) per participant, whereas Qualtrics was free for us to use through our university. Data were also sometimes hard to download from Gorilla compared to Qualtrics, as the servers sometimes became blocked up due to heavy volume.

In terms of setup, both the Qualtrics mental rotation and spatial vocabulary tasks were administered by an experimenter *via* screen share. Because we had to link the participants' baseline and post-test assessments to each other, each child was assigned a random ID number that the experimenter filled out along with other basic information at the beginning of the Qualtrics forms. Parents did not have to fill out anything on Qualtrics for the screen share tasks; all necessary identification information was filled out by the experimenter like in a lab setting. However, since the pattern extension task took place on the child's own touchscreen and not through screen share, the parent was responsible for entering the child's ID number on their own. We found that sending the ID number to parents along with the pattern extension Qualtrics link the night before saved time during the session itself and reduced confusion. We also sent the Gorilla link the night before and instructed parents to complete the first two pages with the consent and demographics information, but not to proceed further. Parents did not have to enter an ID number into Gorilla because our research team was able to enter it from our end before sending out the link. However, parents were asked to manually input information into both platforms at some point, and in spite of the emailed instructions, some had difficulty navigating between both links and remembering which information belonged in which link. To ease the burden on parents, future remote developmental researchers should streamline their methods by limiting themselves to a single platform and reducing the amount of information they have to enter that is typically inputted by experimenters.

As mentioned previously, a notable advantage to remote data collection was our ability to recruit from a wide geographic area, allowing children in areas far from universities to participate in developmental research, an opportunity both we and they would not have had otherwise. We were also able to obtain a larger sample than we have been able to acquire in past in-person training studies ($N = 118$ participants took part in the current study compared to $N = 84$ participants that took part in Casasola et al. (2020)). It should be noted, however, that due to the technological requirements and recruitment methods we used, our sample was not very ethnically or socioeconomically diverse ($n = 87$ participants identified as White/Caucasian, $n = 112$ participants came from middle and upper socioeconomic classes, as defined by maternal education). We also recruited heavily from parenting-based Facebook groups, so the nature of our sample was impacted by the types of families that seek out those types of groups to join. Wider recruitment benefits the field by providing researchers access to samples that are more representative of the general population, and future online research should supplement the inherent geographic diversity of remote research by making a concerted

effort to reach out to online communities with connections to families from a wider variety of ethnic and socioeconomic backgrounds. Online research has the potential to be integrated as a fruitful avenue of recruitment even after the pandemic, although it should be viewed as an addition rather than a substitute for in-person methods, as there are some samples that cannot effectively be reached by remote methods. For example, in addition to technological requirements, online research also requires a sufficiently quiet and spacious home environment that is not available to all families.

Furthermore, if remote studies are to continue even after the pandemic ends, it is essential to verify that the virtual formats of tasks achieve the same internal validity as their in-person counterparts (Scott and Schulz, 2017; Rhodes et al., 2020; Oliver and Pike, 2021). Of the six baseline and post-test assessments in the current study, the mental rotation task and spatial vocabulary assessment matched in age range and format to an in-person study in our lab examining how children's play behaviors shape their spatial skills. Both mental rotation assessments were based on the PRT from Quaiser-Pohl (2003), administered through Qualtrics, used the same number of items, and had the same scoring system. We computed the Cronbach's alpha for both versions of the mental rotation task in the current study (Version A: $\alpha = 0.725$; Version B: $\alpha = 0.713$) and both versions of the mental rotation task in the in-person study (Version A: $\alpha = 0.768$; Version B: $\alpha = 0.723$), which indicated comparable internal reliability across the two tasks. After accounting for the effect of age by calculating residuals, two one-sided *t*-test (TOST) equivalence was calculated for the two versions of the task using the TOSTER package in R (Lakens, 2017). According to this test, we can reject effects larger than $d = 1$, $t(86.32) = 4.893$, $p < 0.0001$, suggesting that the difference between the two task formats is less than one standard deviation from zero. A boxplot depicting the overlap in the residual scores for the two versions of the task can be found in Figure 8. The statistics from the current study, which was the first to examine mental rotation through interactive online methods, produce a promising outlook on the future use of remote methodologies to test spatial-cognitive skills, as they appear to achieve equivalent effects in an online format.

The spatial vocabulary assessment was created by our lab and had been administered during the same in-person study as the mental rotation task. Both assessments were on Qualtrics and contained the same items in the same order. Once again, after accounting for age by calculating residuals, TOST equivalence across the two study formats was calculated for both the expressive and receptive vocabulary portions of the assessment using TOSTER (Lakens, 2017). The results indicated that effects larger than $d = 1$, $t(94.27) = -5.437$, $p < 0.0001$ for expressive vocabulary and larger than $d = 1$, $t(114.07) = 5.786$, $p < 0.0001$ for receptive vocabulary can be rejected, suggesting that difference between task formats for both expressive and receptive vocabulary is less than one standard deviation from zero. Boxplots of the age-adjusted residuals for the two versions can be found in Figure 9 (expressive vocabulary) and Figure 10 (receptive vocabulary). It appears that children achieved similar results on the

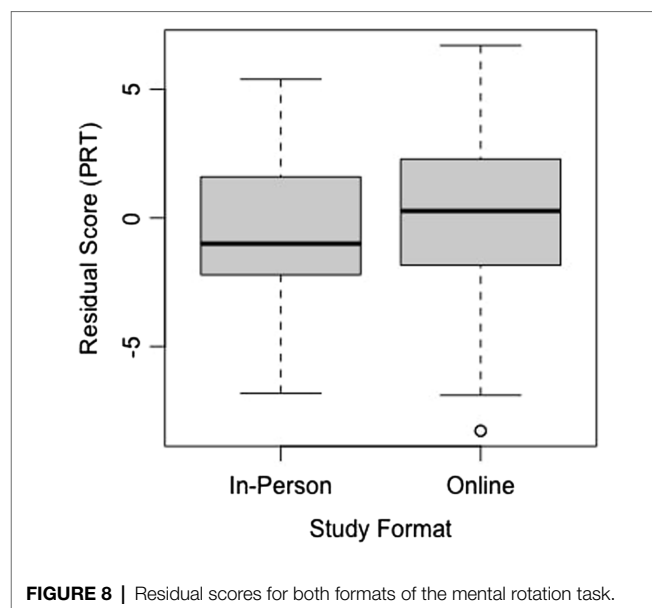


FIGURE 8 | Residual scores for both formats of the mental rotation task.

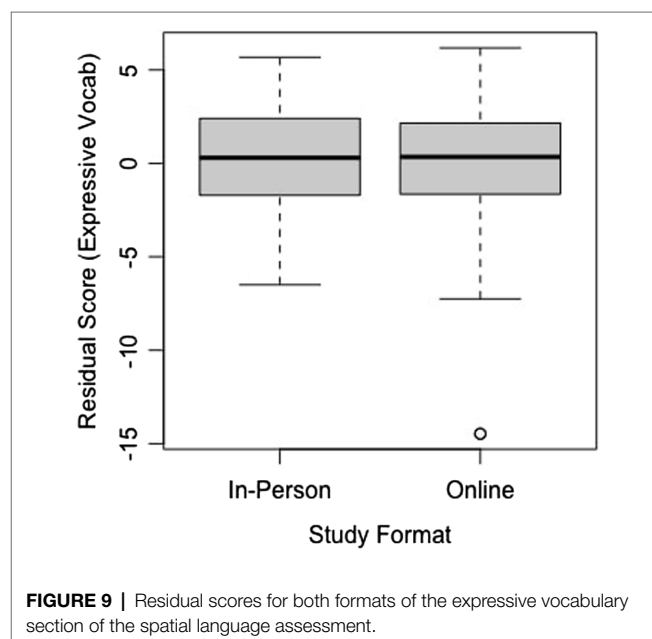
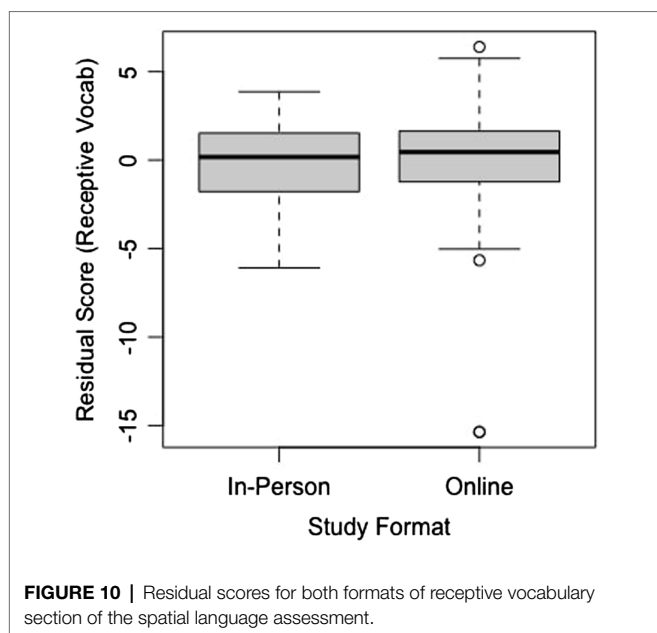


FIGURE 9 | Residual scores for both formats of the expressive vocabulary section of the spatial language assessment.

assessment regardless of whether it took place online or in-person. In line with previous remote methods that have tested language learning and knowledge virtually, this finding indicates that our spatial vocabulary assessment can be used reliably in an online format for this age group.

However, although the TOST equivalence found the two study formats to be statistically equivalent at the inputted parameters, it is noteworthy to mention that there were more extreme residual values (notably at the low end) in the online format, as can be seen from the boxplots. These values may have resulted from children becoming more distracted or less engaged in the online format and thus performing substantially below the mean for



their age group. It is important to keep these possibilities in mind when interpreting data from online studies, because distractibility and engagement may not always be obvious from watching the sessions, making it difficult to successfully exclude every child who lost focus on the task.

Our remaining assessments and the training activities had no comparable in-person task from our lab for comparison (pattern extension task, Gorilla tasks, modified Beery DTVMI, and training activities). Further work is needed to compare the validity and reliability of these tasks when they are conducted in-person as opposed to a remote format, particularly for tasks, such as the modified Beery DTVMI and the drawing activities from the training studies, whose scoring involves a degree of subjectivity.

FINAL TAKEAWAYS AND ADVICE

Entering into the foray of online research, especially with a multi-session training study, introduced our team to unexpected situations that helped us develop an effective protocol for successfully conducting research remotely. As many other researchers in developmental psychology and other fields are approaching the task of adapting their own studies to this new format, we thought it would be helpful to share the more miscellaneous adaptations we employed during our sessions to ensure they proceeded according to plan. Some of this advice is specific to online studies and some would apply to either in-person or online research.

- When emailing links to parents, it is important to let them know when they can open them and how much they can fill out ahead of the session. If you do not want a parent to open a link before the session at all it is best to wait until the session begins to send it.

- Make sure parents are aware when sessions are being audio and video recorded, for what purpose, where the videos will be stored, and who will have access to them.
- It is easy for videos to get washed out, especially if the participant is sitting near a window. We always had our researchers take some time at the beginning of each session to politely ask the parent to adjust the camera until they could see what they needed to see.
- Be aware of how recording works on the platform you are using. For example, when Zoom is set to speaker view it only records video of whoever is speaking at the moment. This feature is disadvantageous when you want a video of the child and not the researcher giving instructions.
- We always had our researchers record on either gallery or spotlight view. When they had to use screen share, we had them expand the video of the participant as large as possible.
- Have your participant use darker colored crayons or markers when they have to physically draw something to ensure that you are able to see what they are drawing. Always have the child hold up whatever they are working on to the screen and make sure it is fully captured by the camera.
- Pay attention to your facial expression and offer consistent encouragement during the session. The child is most likely looking at a close-up of your face the entire time.
- It was helpful to be flexible about task order in our online format. We would recommend it if possible because it helps children maintain attention.
- Be sure to debrief the parent and child (in an age-appropriate way) at the end of the study so they know what the study was about.
- Send compensation as soon as possible after the session.
- Follow-up with parents when you know the results to give them a summary of what you found. This helps them feel included in the research process.

In short, although the widespread shift to online studies was not a voluntary one, with careful planning and study design online studies can provide a valuable source of data for developmental science that augments what researchers are able to accomplish with conventional data collection methods.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Cornell University Institutional Review Board (IRB) protocol number: 1210003363. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

VB and MC conceptualized the idea for the study questions and procedure. VB coordinated all details related to data collection, oversaw coding and data processing, and wrote the manuscript. MC edited the manuscript and obtained the funding used during the study. All authors contributed to the article and approved the submitted version.

FUNDING

A USDA Hatch Grant (219-20-185) was awarded to MC for supporting research on how language promotes children's spatial

skills. This funding was used for participant compensation and to fund two undergraduate research assistants.

ACKNOWLEDGMENTS

We would like to thank Naina Murthy, Kira Lee-Genzel, Kelly Zhou, Stacey Li, Rachel Bank, Nicole Werner, Madeline Hanscom, Michelle Yang, Isabella Lepez, Shayna Morgan, Radiah Khandokar, Talia Petigrow, Chaelin Lee, Brittanie Pic, Marie Hwang, and Mary Simpson, the enthusiastic research assistants who contributed to this project, and all the children and families who dedicated their time to our research in such a busy and uncertain time.

REFERENCES

- Beery, K. E. (2004). *Beery VMI: The Beery-Buktenica Developmental Test of Visual-Motor Integration*. Minneapolis, MN: Pearson.
- Casasola, M., Wei, W. S., Suh, D. D., Donskoy, P., and Ransom, A. (2020). Children's exposure to spatial language promotes their spatial thinking. *J. Exp. Psychol. Gen.* 149, 1116–1136. doi: 10.1037/xge0000699
- Gade, M., Zoelch, C., and Seitz-Stein, K. (2017). Training of visual-spatial working memory in preschool children. *Adv. Cogn. Psychol.* 13, 177–187. doi: 10.5709/acp-0217-7
- Hale, C. M., and Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Dev. Sci.* 6, 346–359. doi: 10.1111/1467-7687.00289
- Hofmann, S. G., Doan, S. N., Sprung, M., Wilson, A., Ebesutani, C., Andrews, L. A., et al. (2016). Training children's theory-of-mind: A meta-analysis of controlled studies. *Cognition* 150, 200–212. doi: 10.1016/j.cognition.2016.01.006
- Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta analyses. *Soc. Psychol. Personal. Sci.* 8, 355–362. doi: 10.1177/1948550617697177
- Malouff, J. M., and Schutte, N. S. (2017). Can psychological interventions increase optimism? A meta-analysis. *J. Posit. Psychol.* 12, 594–604. doi: 10.1080/17439760.2016.1221122
- Mix, K. S., Levine, S. C., Cheng, Y. L., Stockton, J. D., and Bower, C. (2020). Effects of spatial training on mathematics in first and sixth grade children. *J. Educ. Psychol.* 113, 304–314. doi: 10.1037/edu0000494
- Oliver, B. R., and Pike, A. (2021). Introducing a novel online observation of parenting behavior: reliability and validation. *Parenting* 21, 168–183. doi: 10.1080/15295192.2019.1694838
- Quaiser-Pohl, C. (2003). The mental cutting test “schnitte” and the picture rotation test—two new measures to assess spatial ability. *Int. J. Test.* 3, 219–231. doi: 10.1207/S15327574IJT0303_2
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Roseberry, S., Hirsh-Pasek, K., and Golinkoff, R. M. (2014). Skype me! Socially contingent interactions help toddlers learn language. *Child Dev.* 85, 956–970. doi: 10.1111/cdev.12166
- Scott, K., and Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., et al. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychol. Bull.* 139, 352–402. doi: 10.1037/a0028446

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Bambha and Casasola. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Measuring Emerging Number Knowledge in Toddlers

Alex M. Silver^{1*}, Leanne Elliott¹, Emily J. Braham¹, Heather J. Bachman², Elizabeth Votruba-Drzal¹, Catherine S. Tamis-LeMonda³, Natasha Cabrera⁴ and Melissa E. Libertus¹

¹Department of Psychology, Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA, United States, ²Department of Health and Human Development, School of Education, University of Pittsburgh, Pittsburgh, PA, United States, ³Department of Applied Psychology, Steinhardt School of Culture, Education and Human Development, New York University, New York, NY, United States, ⁴Department of Human Development and Quantitative Methodology, University of Maryland, College Park, MD, United States

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Koleen McCrink,
Columbia University, United States
Sara Cordes,
Boston College, United States

*Correspondence:

Alex M. Silver
ams645@pitt.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 April 2021

Accepted: 29 June 2021

Published: 20 July 2021

Citation:

Silver AM, Elliott L, Braham EJ,
Bachman HJ, Votruba-Drzal E,
Tamis-LeMonda CS, Cabrera N and
Libertus ME (2021) Measuring
Emerging Number Knowledge in
Toddlers.
Front. Psychol. 12:703598.
doi: 10.3389/fpsyg.2021.703598

Recent evidence suggests that infants and toddlers may recognize counting as numerically relevant long before they are able to count or understand the cardinal meaning of number words. The Give-N task, which asks children to produce sets of objects in different quantities, is commonly used to test children's cardinal number knowledge and understanding of exact number words but does not capture children's preliminary understanding of number words and is difficult to administer remotely. Here, we asked whether toddlers correctly map number words to the referred quantities in a two-alternative forced choice Point-to-X task (e.g., "Which has three?"). Two- to three-year-old toddlers ($N = 100$) completed a Give-N task and a Point-to-X task through in-person testing or online via videoconferencing software. Across number-word trials in Point-to-X, toddlers pointed to the correct image more often than predicted by chance, indicating that they had some understanding of the prompted number word that allowed them to rule out incorrect responses, despite limited understanding of exact cardinal values. No differences in Point-to-X performance were seen for children tested in-person versus remotely. Children with better understanding of exact number words as indicated on the Give-N task also answered more trials correctly in Point-to-X. Critically, in-depth analyses of Point-to-X performance for children who were identified as 1- or 2-knowers on Give-N showed that 1-knowers do not show a preliminary understanding of numbers above their knower-level, whereas 2-knowers do. As researchers move to administering assessments remotely, the Point-to-X task promises to be an easy-to-administer alternative to Give-N for measuring children's emerging number knowledge and capturing nuances in children's number-word knowledge that Give-N may miss.

Keywords: number knowledge, math development, cardinal principle, remote data collection, toddlers aged 12 to 36 months

INTRODUCTION

Individual differences in math relate to academic achievement, career choice, employment and income, and health and financial decision-making (e.g., Trusty et al., 2000; Currie and Thomas, 2001; Duncan et al., 2007; Reyna and Brainerd, 2007; Agarwal and Mazumder, 2013). Critically, large variability in math performance is present among children even at the start of formal education (Jordan et al., 2006). Much work has attempted to understand the development of early numerical skills in the hope of understanding sources of early emerging individual differences.

When examining numerical skills, even at young ages, it is critical to consider the distinct skills that fall under this domain. Research suggests that from birth, humans possess the ability to discriminate and precisely represent small numbers of objects *via* the object-file system and imprecisely represent larger quantities *via* the approximate number system (ANS; see Feigenson et al., 2004). Non-symbolic number representations in the object-file system are precise but limited to only a few items (typically 1, 2, and 3 in infants and toddlers), whereas representations in the ANS are imprecise but extend to larger quantities (4+). As such, discrimination of two quantities using the ANS is ratio dependent, such that it is easier to discriminate between quantities that have a larger relative difference (i.e., 6 vs. 12 or 12 vs. 24 objects) than quantities that are closer together (i.e., 6 vs. 9 or 12 vs. 18 objects; Dehaene et al., 1998; Libertus and Brannon, 2009).

These non-symbolic number systems are often contrasted with the symbolic number system, in which number words and other symbols map to their exact quantities. Previous work suggests that children come to understand the meaning of exact number words very slowly (Wynn, 1990, 1992): English-speaking children first learn the meaning of the word “one” around two-and-a-half years of age but lack knowledge of numbers larger than one. About four to five months after learning the meaning of “one,” children understand the word “two” but not larger numbers, such as “three” or “four.” It takes several more months for children to display knowledge of the word “three.” Children who display knowledge of some but not all number words are typically referred to as “subset knowers” (Le Corre and Carey, 2007). Not until children are three or four years of age do they fully grasp the cardinality principle—that each number word refers only to an exact set of that quantity with the last number in the count list referring to the total number of items in the set (see Carey, 2009, for review).

This estimated timeline indicates the ages at which children have a complete understanding of each number word and can successfully create sets of that quantity. Although infants and toddlers may not fully understand the meaning of number words, recent work suggests they show an early sensitivity to counting. Eighteen-month-old infants showed a preference for correctly ordered counting sequences; that is, although they were unable to recite the count list themselves, they recognized and preferred to listen to the correct order of the number words (Ip et al., 2018). Similarly, 14- to 18-month-old infants

appear to be able to use their ability to recognize the count list to help them overcome typical memory limits (Wang and Feigenson, 2019). Infants generally display working memory capacity limits of three items and fail to remember the number of hidden items when it exceeds this limit (Feigenson and Carey, 2003). However, when objects are counted before being hidden, infants are able to overcome this memory limit (Wang and Feigenson, 2019). Thus, even though infants may not grasp the full meaning of number words, they may still be aware of the numerical nature of these words and may be able to use this knowledge despite lacking precise representations of the quantities.

Other studies with toddlers and preschool-aged children also suggest that young children have preliminary, noisy understandings of number words prior to developing more precise mappings between the words and the quantities to which they refer (Wagner et al., 2019; O’Rear et al., 2020). Specifically, before learning the exact meanings for small numbers, two- to five-year-old children display some preliminary knowledge of those number words and are able to create sets of that size more often than predicted by chance (Wagner et al., 2019). Similarly, three- to five-year-old children who did not fully understand a number word nevertheless still displayed some partial knowledge when asked to produce a set of that size, and this partial knowledge predicted their likelihood of fully understanding that number word a few weeks later (O’Rear et al., 2020). Together, these studies suggest that young children have an early recognition of number words that they may use to then refine their understanding of numbers.

Measuring Number Knowledge

Acquisition of number-word meanings is typically measured using the “Give-a-Number” task (i.e., Give-N). Give-N assesses children’s understanding of exact number words (Wynn, 1990, 1992). Children are required to produce sets of objects in various quantities (e.g., “Can you give me three fish?”), with the highest number they can correctly and reliably produce in a set defining their “knower-level.” However, by grouping children into discrete knower-level categories, Give-N may not capture approximate knowledge of number words, that is, children’s preliminary understanding of number words prior to understanding the exact meaning of a number word (Wagner et al., 2019; O’Rear et al., 2020). Furthermore, the Give-N task may place high demands on working memory and attention, because children must hold in memory the number of items they are supposed to generate as they attend to counting out the set, which may underestimate children’s true number knowledge (see Frye et al., 1989; Cordes and Gelman, 2005; but see Le Corre et al., 2006). Additionally, Give-N requires physical materials for administration which may be difficult to standardize and supply to participants in studies requiring remote administration.

The Point-to-X task (see Wynn, 1992; Levine et al., 2010; Gunderson and Levine, 2011; van Marle et al., 2014; O’Rear et al., 2020) offers an alternative approach to assessing children’s number knowledge. Point-to-X is a forced-choice response task in which researchers present children with two images and

prompt them to select one by pointing (i.e., “Which has three?”). The two images typically display sets of objects that differ only in number. Previous versions of this task asked children to compare adjacent numbers (one-away; Wynn, 1992); used a limited number range from 1 to 6 (Wynn, 1992; Levine et al., 2010; Gunderson and Levine, 2011; O’Rear et al., 2020); tended to focus on either exclusively small or large number response options in a given trial (van Marle et al., 2014); did not include specified practice trials to introduce participants to the task (Levine et al., 2010; van Marle et al., 2014); or used practice trials that included numbers with no control for children’s general ability to follow directions (Gunderson and Levine, 2011; O’Rear et al., 2020). As a result, it was not always possible to test for approximate understanding of the involved numbers if they were very close together, test for comparisons of larger numbers or between small and large numbers, or control for children’s general ability to follow directions in the task.

Finally, previous studies of Point-to-X were conducted solely in-person, so whether this task can be successfully administered remotely remains an open question. Given the recent transition to remote data collection in the field in large part fueled by the COVID-19 pandemic, validating procedures that could be utilized both in-person and remotely is a crucial step. Importantly, remote data collection holds the potential to test participants who otherwise may not be able or may be highly unlikely to participate in research studies. Thus, the need to compare in-person and remote data collection methods transcends the current pandemic-related needs and will hopefully pave the way to test more representative samples in our research in the future.

The Current Study

We developed a novel version of Point-to-X to assess children’s number knowledge and expand on the types of comparisons used in prior versions of the task. Specifically, we included a larger range of numbers, more varied types of number comparisons, word-control practice trials to control for children’s general ability to follow directions, and a procedure for both in-person and remote administration. We compared children’s performance in this novel Point-to-X task to performance in a traditional Give-N task to probe whether we can capture nuances in their number knowledge better by grouping children into discrete knower-levels of Give-N.

We had three aims. First, we aimed to identify whether this novel Point-to-X task accurately tapped toddlers’ number knowledge when comparing performance to chance, and to validate the use of the novel Point-to-X measure for in-person and online data collection. Second, we explored whether children’s performance differs on different trial types of the Point-to-X task (e.g., trials where the options differ in distance, target size, or response option size). Finally, we aimed to compare performance in the Point-to-X task to a traditional Give-N task and explore children’s performance on Point-to-X trials above their Give-N knower-level.

To identify whether the Point-to-X task taps children’s number knowledge, we compared performance to chance and compared

performance for children tested in-person and those tested remotely. Based on work studying the ANS in young children (e.g., Halberda and Feigenson, 2008; Navarro et al., 2018), we expected that toddlers would show greater performance on trials where the response options were far away from each other (i.e., there was a larger ratio between the two quantities, such as a comparison between 4 and 10) compared to trials where the options were only one or two away (i.e., the ratio between the two quantities was much smaller and thus harder to discriminate, such as comparisons between 4 and 5 or 4 and 6). Furthermore, we predicted that children would perform better on trials where the requested target number was small (closer to children’s knowledge level) than on trials where the target was large, and similarly, that children’s performance would be better on trials where the numbers were both small (and thus closer to children’s knowledge level). Finally, we predicted that children’s performance in the novel Point-to-X task would positively, yet only moderately, correlate with their performance on a Give-N task (see O’Rear et al., 2020), as we expected to find greater individual variability in the Point-to-X task than Give-N. To probe children’s number knowledge in more detail, we explored whether children at various knower-levels may perform above chance on Point-to-X trials above their knowledge level. Based on recent work suggesting children may display partial knowledge of number words before fully understanding their meanings (e.g., Wagner et al., 2019; O’Rear et al., 2020), we expected that children would perform above chance, even on trials containing numbers above their knower-level.

MATERIALS AND METHODS

Participants

Participants were 100 toddlers (56 girls) ranging in age from 2 years 1 month to 3 years 2 months (child *M* age = 2 years 8 months, *SD* = 2.8 months). Thirty-three children were tested in-person and 67 children remotely. Children were reported by their parents to be predominantly White, non-Hispanic (64%); 12% were White, Hispanic/Latino; 9% were Black/African-American, non-Hispanic; 1% were Asian, non-Hispanic; 7% were multi-ethnic, and 7% did not have their race and ethnicity reported. Children were tested in their preferred language (English or Spanish), with 92% of children tested in English.

An additional 59 children participated but were dropped from analyses due to refusal to attempt the Point-to-X task (11), refusal to complete the Point-to-X task after starting (17), experimenter error in the Point-to-X task (2), use of the stopping rule in the Point-to-X task (13), or exclusion for incorrect responses on the practice trials of the Point-to-X task (16). We compared children excluded from analyses to those included to identify if data were missing at random or instead showed systematic patterns of missingness. Children excluded from analyses did not differ from those included in analyses in age, $\chi^2(132) = 140.80$, $p = 0.284$, or type of testing (26 in-person vs. 33 remote excluded), $\chi^2(1) = 1.72$, $p = 0.163$.

Children excluded from analyses were more likely to be boys (31 boys excluded), $\chi^2(1) = 4.88$, $p = 0.027$, and more likely to be tested in Spanish (15 Spanish-tested excluded), $\chi^2(1) = 9.10$, $p = 0.003$. However, these latter results should be treated with caution due to the small number of children tested in Spanish.

All parents were instructed not to interact or provide encouragement to their children, or otherwise react to children's responses. They were reminded of this rule before each task. For trials where parents interfered after children had already made a response, we coded children's initial response as their final choice. For trials where parents interfered before children responded, we excluded children's responses for those trials.

Procedure

Families were recruited from three cities in the United States (all mid-Atlantic metropolitan areas) through a combination of flyers, online postings, and mailings, and were compensated \$50 for their time. They were told that the study was designed to study how parents support their children's early learning but were not told about the focus on math. Prior to data collection, parents provided written informed consent as approved by the local Institutional Review Boards. Data are drawn from testing of children during an in-person home visit ($n = 33$; April 2019–March 2020 before the COVID-19 lockdown) or on a Zoom video call ($n = 67$; post-July 2020). Children completed a Point-to-X task and a Give-N task. Assessments were video recorded (*via* either video cameras in-person or Zoom video recording) and coded by trained researchers. In addition to the measures included in the current analyses (described below), children completed assessments of their non-symbolic numerical comparison abilities and spatial knowledge and their parents completed math assessments, questionnaires about their family, and participated in semi-structured observations with their children as part of the larger study. These measures were not in the focus of the current paper and thus are not discussed further.

Most children ($n = 91$) completed the Give-N task first. There was no difference in children's performance in the Point-to-X task or the Give-N task based on the order of task administration, $\chi^2(9) = 9.52$, $p = 0.391$, and $\chi^2(6) = 2.26$, $p = 0.894$, respectively.

Measures

Point-to-X

A novel Point-to-X task was created for this study (see **Appendix** for items). Children tested in-person in their homes viewed a series of images printed on individual sheets of laminated paper presented by the experimenter on each trial. Children tested remotely were mailed a set of the paper materials in a binder prior to the session, and the experimenter administered the verbal prompts *via* Zoom as parents turned the pages for each trial.

All children, regardless of method of testing (in-person or remote), received the same set of Point-to-X items. To familiarize children with the Point-to-X task, children were first given two practice trials with different common objects and were

prompted to point to one image (e.g., "Which has a ball?"). Subsequently, in twelve number-word trials, each image showed two sets of identical stimuli differing only in number (e.g., four ducks and five ducks), and children were prompted to point to one of the images (e.g., "Which has four ducks?"). Number-word trials varied along three distinct dimensions: (1) the numerical distance between the two sets [for "one-away" trials, the numbers differed by one; for "two-away" trials, the numbers differed by two; and for "far-away" trials, the numbers differed by more than four]; (2) the size of the target number [for eight trials, the prompted number was small (1–4), and for four trials, the prompted number was large (5–10)]; and (3) the size of the response options [for five trials, both numbers were small (1–4), and for seven trials, at least one number was large (5–10)]. The side of the correct response was counterbalanced across trials.

When administering the task, if children initially pointed to one image, then verbally indicated that they wanted to change their answer, the second point was counted as their response. In cases where children did not respond, the experimenter repeated the prompt one time. If children still did not respond, the experimenter moved on to the next trial and children received zero points for the trial. If children pointed to both images without clearly signaling which was their preferred response, the experimenter prompted, "Remember, you can only choose one. Which has [number]?" After this prompt, if children continued to point to both images, they received zero points for the trial. If children responded incorrectly to each of the first three number-word trials, the experimenter employed a stopping rule and ended the task. Task duration for children included in analyses ranged from 1:50 to 8:45 min, with an average of 4:29 min ($SD = 1:31$).

Videos were coded by trained researchers who identified the image children pointed to for each trial. Children received one point for pointing to the correct image, or zero points for pointing to the incorrect image. 30% of videos (47 out of 159) were double-coded by a second researcher to assess inter-coder reliability. Coders agreed for 98.2% of trials. Disagreements were resolved by a third coder. Children's Point-to-X score is the percentage of trials that contained correct points.

Give-N

Children's knower-level was assessed using a modified Give-N task (Wynn, 1990, 1992). Children tested remotely were sent a set of the materials (a plate and 10 plastic objects) prior to the testing session, and the experimenter administered the verbal prompts with the puppet *via* Zoom as children's parents helped facilitate the clearing of the plate after each trial.

Children were shown an animal puppet held up by the experimenter and a large pile of plastic objects that could be considered food (e.g., peanuts and fish). To introduce children to the game, children were shown the puppet and told that the puppet loves to eat snacks. They were asked to help "feed" the puppet by putting out the correct number of objects for the puppet to eat (either in front of the puppet for children tested in-person or on the plate for children tested remotely). The experimenter then said "Look, let us feed [name of puppet]!"

and mimed placing an object from the large pile in front of the child in a new pile in front of the puppet (in-person) or mimed placing an object on a plate that the experimenter held (for children tested remotely). Then, the experimenter held the puppet up to the object (in-person) or the webcam (remotely) and enacted the puppet “eating” the objects and saying, “Yum yum yum!”

Once the practice trial was completed, test trials began. The researcher asked children to “feed” the puppet different numbers of objects by placing the objects in a pile. For each trial, children were asked “Can you give [name of puppet] [number] [name of food]?” and instructed to put the set of objects in a new pile for the puppet to eat. After the child paused for more than 3 s or indicated that they were done creating the set, the experimenter prompted confirmation from the children, “Is that [number]?” If children said yes or nodded, the experimenter held the puppet up to the pile (in-person) or the webcam (remotely) and said, “Yum yum yum! Thank you!” If children said no or shook their head, they were given one chance to correct their response and were instructed, “Ok, well [name of puppet] wants [number] [name of food]. Can you give [name of puppet] [number] [name of food]?” Once children had adjusted the number of objects or paused for more than 3 s, the experimenter held the puppet up to the pile of objects (in-person) or the webcam (remotely) and said “Yum yum yum! Thank you!” The objects were then returned to the main pile before the next trial. If children did not respond to a trial, the experimenter repeated the prompt one time. If children still did not respond, the experimenter moved on to the next trial and children were considered to have responded incorrectly and received zero points for that trial.

Trials were administered in a titrated manner (see Wynn, 1990, 1992). All children were first asked for one object and then for two objects. If a child correctly responded to a trial, they were then tested with the next number in the sequence (e.g., asked for three after responding correctly to two). If a child responded incorrectly to a trial, they were subsequently asked for the next smaller number (e.g., asked for one after responding incorrectly to two). This process was repeated until children successfully produced a set of N objects twice and failed to produce $N+1$ twice. Task duration ranged from 1:05 to 10:35 min, with an average of 3:12 min ($SD = 1:43$).

After administration, videos were coded by trained researchers who credited children with one point for each set of the correct number of objects. 70% of videos (112 out of 159) were double-coded by a second researcher to ensure reliability. Coders agreed for 89.5% of “knower-level” scores. Any disagreements were resolved by a third coder. Children were not given any feedback on their performance, and the highest number at which they produced the correct set size twice while failing twice at the next highest number was used here as their Give-N “knower-level” score. As a robustness check, we also calculated children’s knower-level score as the highest number at which they produced the correct set size twice and *did not produce that set size for any other number* (e.g., to be classified as a 2-knower they successfully produced 2 objects when asked for two and did not produce 2 objects when asked for any other

number), but using this stricter criterion for knower-level did yield differences in the pattern of results. Thus, analyses are based on the highest number that children correctly produced twice as their Give-N knower-level score.

Analysis Plan

All analyses were conducted using Stata/SE 15.1 (StataCorp, 2017). We first examined descriptive statistics for children’s overall performance in the Point-to-X task. To test whether children’s performance in the Point-to-X task was significantly above chance, we used a one-sample t -test comparing the mean performance across all trials to 50% (i.e., expected performance if children were simply guessing for each trial). We then examined whether children’s performance in Point-to-X was related to children’s age using a pairwise correlation and whether performance differed based on children’s sex or mode of testing using one-way ANOVAs. Additionally, we tested whether children’s age differentially related to their performance on Point-to-X based on whether they were tested in-person vs. remotely using a linear regression model with main effects of children’s age and mode of testing and an interaction term between them.

We next examined children’s performance on Point-to-X trial subtypes, and whether performance on each subtype differentially related to children’s age using tests of equality of the correlation coefficients. We also tested whether performance in each of the trial subtypes differed based on whether they were tested in-person vs. remotely using one-way ANOVAs.

Then, we asked whether children’s performance in the Point-to-X task differed for trials of different numerical distances. We compared the mean performance for one-away trials, two-away trials, and far-away trials using a one-sample multivariate test on the means. Similarly, we used a paired t -test to address whether children’s performance in the Point-to-X task differed for trials where the target number was small (i.e., the number asked for was between 1 and 4) vs. trials where the target number was large (i.e., the number asked for was between 5 and 10). We then addressed whether children’s performance in the Point-to-X task differed for trials where both response options were small (between 1 and 4) vs. trials where at least one option was large (between 5 and 10) using a paired t -test, although we note that for the former, these trials were all fairly close comparisons. To control for the distance between options in these comparisons, we also examined performance using paired t -tests on trials where response options were both small and differed by one to trials where the response options included at least one large number and differed by one. We similarly compared performance on trials where response options were both small and differed by two to trials where the response options included at least one large number and differed by two.

Finally, we turned to examining children’s performance on the Give-N measure. Using a Pearson’s chi-squared test, we examined whether children’s Give-N performance differed based on whether they were tested in-person or remotely. We examined how performance in the Point-to-X task related to children’s performance in the traditional Give-N assessment

by performing a one-way ANOVA of Point-to-X performance using children's Give-N knower-level score as the factor variable as well as by calculating a pairwise correlation between children's Point-to-X performance score and their Give-N knower-level score. To control for child age, we calculated a partial correlation between children's Point-to-X performance and their Give-N knower-level score that covaried any effects of age. We then examined whether the relation between performance on Point-to-X and children's Give-N knower-level differed based on whether they were tested in-person vs. remotely by using a linear regression model with main effects of Give-N knower-level and mode of testing and an interaction term between them.

In addition, we performed detailed analyses of children's performance in Point-to-X as a function of their knower-level scores. Specifically, to determine whether Point-to-X is sensitive to an approximate understanding of number words, we compared all children's performance on trials in the Point-to-X task that were within their knower-level and those outside of their knower-level to chance using one-sample *t*-tests. We also looked at these trials specifically for 1-knowers and 2-knowers, the largest two groups of subset-knowers in our sample, as well as a 3-knowers and 4-knowers combined together due to small group sizes, to identify possible differences in their approximate understanding of number words. Given recent work suggesting that children have preliminary understandings of numbers above their knower-level, but only for small sets (Wagner et al., 2019), we also compared performance on trials outside children's knower-level that contain only small number response options to chance using one-sample *t*-tests.

RESULTS

Overall Performance in Point-to-X

Descriptive statistics for children's performance on each trial of the Point-to-X task are presented in **Table 1**. Performance did not differ for children tested in-person vs. remotely ($p = 0.142$). Across all trials, performance in the Point-to-X task averaged 65.25% correct, which differed significantly from chance responding, $t(99) = 8.80$, $p < 0.0001$. Sixty-nine percent of children scored above chance on the task. Performance did not differ based on children's sex ($p = 0.469$). However, children's age predicted performance in the Point-to-X task, such that a 1 *SD* increase in children's age in months was associated with a 0.27 *SD* increase in children's performance on the task ($p = 0.007$). The mode of testing did not moderate the association between children's age and their Point-to-X performance ($\beta = 0.09$, $p = 0.600$). Children's age did not differentially relate to performance in any of the trial subtypes examined (all $ps > 0.265$), and so we did not include age as a factor in further analyses.

Performance in Trial Subtypes of Point-to-X

Descriptive statistics for children's performance in different trial types of the Point-to-X task are presented in **Table 2**. Notably, performance did not differ for children tested in-person vs.

those tested remotely for any of the trial subtypes examined (all $ps > 0.05$). We first examined children's performance for trials of different distances. Specifically, we tested whether children differed in performance on trials where response options were one-away, two-away, or far-away. Contrary to hypotheses, children did not differ on their performance for one-away, two-away, or far-away trials, Hotelling $F(2,98) = 0.37$, $p = 0.692$.

We next examined whether children's performance differed for trials where the target number was small vs. trials where the target number was large. Although performance was higher for trials where the target number was small ($M = 67.25\%$, $SD = 22.03\%$) vs. large ($M = 61.25\%$, $SD = 27.15\%$), the difference was only marginally significant, $t(99) = 1.72$, $p = 0.088$.

However, children's performance differed for trials where the response options were both small vs. trials where at least one of the response options was a large number. Specifically, as hypothesized, performance was significantly better for trials where both response options were small, $t(99) = 3.53$, $p < 0.001$. Because the distance between options when both response options were small could not be far-away (i.e., the options ranged from 1 to 4 and thus could not be more than 3 apart), we compared performance on trials where response options were both small and differed by one to trials where the response options were not both small and differed by one, to control the distance. We found that performance was significantly better for trials where both response options were small, $t(99) = 2.91$, $p = 0.004$. Similarly, we compared performance on trials where the response options were both small and differed by two to trials where the response options were not both small and differed by two, to control the distance. Again, performance was significantly better for trials where both response options were small, $t(99) = 3.92$, $p < 0.001$. Thus, children's performance was significantly better for trials where both response options were small even when the distance between numbers was held constant.

Relations Between Point-to-X Performance and Give-N Performance

Our final aim was to compare children's performance on the Point-to-X task with their performance on a traditional Give-N task. Of the 100 children included in analyses of the Point-to-X task, 15 did not have usable data from the Give-N task due to refusal to complete the task (7), the task not being administered by the experimenter (1), or experimenter error while administering the task (7). As such, we examined how children's Give-N knower-level score was related to their Point-to-X score for the remaining 85 children.

Children's Give-N knower-levels ranged from 0-knowers to 6-knowers in this sample (**Table 3**). Give-N performance did not differ for children tested in-person versus remotely ($p = 0.285$). A one-way ANOVA indicated that performance in the Point-to-X task significantly differed based on children's Give-N knower-level score, $F(6,78) = 11.31$, $p < 0.001$. Furthermore, higher scores in the Point-to-X task were associated with higher Give-N knower-level scores, $r = 0.64$, $p < 0.001$.

TABLE 1 | Descriptive statistics for children's performance in the Point-to-X task, $N = 100$.

| Trial | Distance | Target size | Options size | <i>M</i> | <i>SD</i> | Different from chance? |
|-------|----------|-------------|--------------------|----------|-----------|------------------------|
| 1 | Two-away | Small | Both small | 81.00 | 39.43 | $t(99) = 7.86^{****}$ |
| 2 | Far-away | Small | At least one large | 74.00 | 44.08 | $t(99) = 5.44^{****}$ |
| 3 | One-away | Small | At least one large | 50.00 | 50.25 | $t(99) = 0.00$ |
| 4 | Far-away | Large | At least one large | 71.00 | 45.60 | $t(99) = 4.60^{****}$ |
| 5 | Two-away | Small | Both small | 68.00 | 46.88 | $t(99) = 3.84^{***}$ |
| 6 | Far-away | Small | At least one large | 60.00 | 49.24 | $t(99) = 2.03^*$ |
| 7 | One-away | Small | Both small | 56.00 | 49.89 | $t(99) = 1.20$ |
| 8 | Two-away | Large | At least one large | 58.00 | 49.60 | $t(99) = 1.61$ |
| 9 | Far-away | Large | At least one large | 60.00 | 49.24 | $t(99) = 2.03^*$ |
| 10 | One-away | Small | Both small | 68.00 | 46.88 | $t(99) = 3.84^{***}$ |
| 11 | Two-away | Large | At least one large | 56.00 | 49.49 | $t(99) = 1.20$ |
| 12 | One-away | Small | Both small | 81.00 | 39.43 | $t(99) = 7.86^{****}$ |

* $p < 0.05$; *** $p < 0.001$ and **** $p < 0.0001$.

TABLE 2 | Descriptive statistics for children's performance in the Point-to-X task, $N = 100$.

| Trial type (Number of trials) | <i>M</i> | <i>SD</i> | Min | Max | Different from chance? |
|----------------------------------|----------|-----------|-----|-----|------------------------|
| All trials (12) | 65.25 | 17.33 | 25 | 100 | $t(99) = 8.80^{****}$ |
| One-away trials (4) | 63.75 | 27.15 | 0 | 100 | $t(99) = 5.06^{****}$ |
| Two-away trials (4) | 65.75 | 28.79 | 0 | 100 | $t(99) = 5.47^{****}$ |
| Far-away trials (4) | 66.25 | 25.22 | 0 | 100 | $t(99) = 6.44^{****}$ |
| Target number is small (8) | 67.25 | 22.03 | 25 | 100 | $t(99) = 7.83^{****}$ |
| Target number is large (4) | 61.25 | 27.15 | 0 | 100 | $t(99) = 4.14^{***}$ |
| Both options are small (5) | 70.80 | 24.02 | 0 | 100 | $t(99) = 8.66^{****}$ |
| At least one option is large (7) | 61.29 | 20.13 | 0 | 100 | $t(99) = 5.61^{****}$ |

*** $p < 0.001$ and **** $p < 0.0001$.

TABLE 3 | Descriptive statistics for children's performance in the Give-N task, $N = 85$.

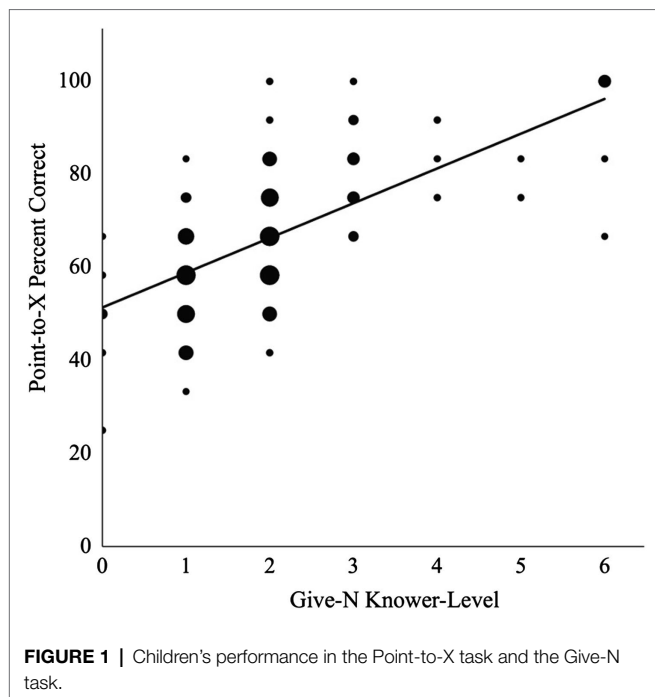
| Knower-level | Number of children | <i>M</i> (<i>SD</i>) Point-to-X score |
|--------------|--------------------|---|
| 0-Knower | 6 | 48.61(14.35) |
| 1-Knower | 26 | 56.73(12.02) |
| 2-Knower | 31 | 67.47(13.50) |
| 3-Knower | 12 | 80.56(10.26) |
| 4-Knower | 3 | 83.33(8.33) |
| 5-Knower | 2 | 79.17(5.89) |
| 6-Knower | 5 | 90.00(14.91) |

This correlation is displayed in **Figure 1**. The partial correlation between performance in Point-to-X and Give-N knower-level scores, when controlling for the contribution of age, remained strong, $r = 0.62$, $p < 0.001$. Furthermore, mode of testing did not moderate the association between children's Give-N knower-level scores and their Point-to-X performance ($\beta = -0.33$, $p = 0.106$). That is, associations between Point-to-X and Give-N were similar for children tested in-person, $r = 0.64$, $p < 0.001$, and remotely, $r = 0.65$, $p < 0.001$.

We then examined children's performance on the Point-to-X task in more detail based on their knower-level. We first looked at trials in the Point-to-X task that were within children's knower-level (e.g., for a 1-knower, trials that included "one" as an option; for a 2-knower, trials that included either "one" or "two"). This analysis excluded 0-knowers ($n = 6$), since

there were no numbers within their knower-level. We found that children's performance on trials including at least one number within their knowledge ($M = 76.87\%$, $SD = 20.58$) was significantly above chance, $t(77) = 11.53$, $p < 0.001$. We next looked at performance on trials in the Point-to-X task that included any numbers above children's knower-level (e.g., for a 1-knower, trials where the smallest number present was any number larger than "one"; for a 2-knower, trials where the smallest number present was any number larger than "two"). We found that children's performance on trials including numbers above their knower-level ($M = 56.76\%$, $SD = 21.38$) was also significantly above chance, $t(75) = 2.76$, $p = 0.007$. We next compared children's performance on trials that were within children's knower-level to performance on trials that were above children's knower-level and found that performance on trials within children's knower-level was significantly better than performance on trials above children's knower-level, $t(69) = 5.29$, $p < 0.001$.

Finally, we compared performance on these types of trials for the two largest groups of subset-knowers: 1-knowers ($n = 26$) and 2-knowers ($n = 31$), as well as a combined group of 3-knowers and 4-knowers ($n = 15$). We found that all of these subset-knowers were significantly above chance for trials that included at least one number within their knowledge ($M_s > 67.95\%$, $p_s < 0.002$). However, for trials where the smallest number was above children's knowledge, 1-knowers did not perform above chance [$M = 53.42\%$, $SD = 12.97$;



$t(25) = 1.34$, $p = 0.191$], whereas 2-knowers performed significantly above chance [$M = 57.47\%$, $SD = 17.59$; $t(28) = 2.29$, $p = 0.030$], and 3-knowers and 4-knowers performed well above 50%, but not statistically significantly due to the small sample size [$M = 64.44\%$, $SD = 36.66$; $t(14) = 1.53$, $p = 0.149$]. Nonetheless, 1-knowers performed significantly above chance for trials where the smallest number was anything above children's knowledge and both response options were small numbers [$M = 61.54\%$, $SD = 22.49$; $t(25) = 2.62$, $p = 0.015$], replicating Wagner et al. (2019).

DISCUSSION

Accurately measuring early math skills has major educational implications, as individual differences in early math performance predict long-term outcomes (e.g., Duncan et al., 2007) and there is a need to accurately identify children who may benefit from early intervention. Typical methods for assessing toddlers' number knowledge provide useful starting points but also highlight the need for development of more nuanced measures. Previous Point-to-X tasks typically only used a limited range of smaller numbers (Wynn, 1992; Levine et al., 2010; Gunderson and Levine, 2011; O'Rear et al., 2020), limited stimuli to closely spaced numbers (Wynn, 1992), and did not always include practice trials to ensure that children understood the task (Levine et al., 2010; van Marle et al., 2014). Meanwhile, the Give-N task may put unnecessary demands on children's cognitive abilities (see Frye et al., 1989; Cordes and Gelman, 2005; but see Le Corre et al., 2006) and may miss important nuances in children's knowledge (see Wagner et al., 2019; O'Rear et al., 2020). Additionally and critically given the recent transition to remote data collection in the field, Give-N

may not be easy to administer remotely due to the required presence of large sets of identical items. Here, we sent materials to families to administer Give-N remotely, but this may not be feasible for many studies and research groups, given the time and financial costs to delivery. Furthermore, sending materials to families is fairly impractical, because scheduling testing visits depends on the timely arrival of those necessary materials and materials not getting lost in the mail or in families' homes.

Our new task expands on previous versions of Point-to-X by including a larger range of numbers, more varied types of number comparisons, and word-control practice trials, with the added aim of administration ease in-person and remotely. Toddlers' performance in the Point-to-X task was significantly above chance for all trial types, suggesting that toddlers have some understanding of the prompted number word that allowed them to rule out incorrect responses, despite their limited understanding of exact cardinal values. Even for trials well beyond their knowledge level, toddlers were able to successfully map the prompted number word to the correct image more often than would be seen if they had simply guessed.

Somewhat surprisingly, children performed equivalently on trials regardless of the distance between response options. This counters our hypotheses that children would be better at selecting the correct option when the response options were farther apart than when they were closer together as we had expected that performance in this task would show the ratio-dependent performance of the ANS. Perhaps, for the far-away trials used here (7 vs. 2, 5 vs. 1, 10 vs. 3, and 4 vs. 10), the ANS was not recruited due to the fact that one of the numbers was always small and the ANS typically is only recruited for comparison of large sets.

On the other hand, children's performance was significantly above chance on all four far-away trials, whereas their performance was only above chance for two of the one-away trials and two of the two-away trials. High performance on these two trials of each type led the overall average for those trial types to be similar to the far-away trials. This high performance was found primarily for trials, including small numbers, whereas performance on one-away and two-away trials, including larger numbers, were only at chance, suggesting an interaction between distance and number size. Unfortunately, we cannot address this possibility because all of the far-away trials included at least one large number due to the criterion of being at least four apart.

Children were best at discriminating small numbers, performing marginally better when the target number was small, and significantly better when both response options were small numbers. Perhaps, children may have more precise representations and partial knowledge of small number words (Wagner et al., 2019; O'Rear et al., 2020). Additionally, children may simply have more exposure to small numbers and thus be more comfortable recognizing them. Indeed, parents are much more likely to talk about small numbers than large numbers with their children (e.g., Dehaene and Mehler, 1992; Elliott et al., 2017).

Furthermore, as hypothesized, toddlers' performance in Point-to-X closely related to their Give-N knower-level, indicating that Point-to-X performance reliably taps children's understanding of exact number words overall. Notably, however, children at a particular Give-N knower-level varied in their Point-to-X performance, suggesting that Point-to-X may capture important individual differences that are missed by grouping children into distinct knower-levels. Importantly, 1-knowers performed significantly above chance on Point-to-X trials including "one" as an option and on trials including only small numbers larger than one as an option, but performed at chance on trials including larger numbers. In contrast, 2-knowers performed significantly above chance on Point-to-X trials including an option within their knower-level (i.e., "one" and "two") and on trials that included numbers above their knower-level. These findings suggest that 2-knowers have a fuller grasp of numbers than do 1-knowers and should not be simply characterized as understanding one additional number word (i.e., "two"). This intriguing finding supports the idea that children's acquisition of the meaning of "one" may be significantly scaffolded by the distinction between singular and plural in the English language (Barner, 2012, 2017) but not distinctions beyond that. An exciting future direction would be to use the Point-to-X task with children learning languages that use dual markings (e.g., Slovenian and Saudi Arabic) to see whether these children learn the meaning of "two" faster (Almoammer et al., 2013) and show an understanding of the approximate meaning of number words above "one" as 1-knowers.

Our findings add to a growing literature suggesting that children have knowledge of number words outside of their knower-level (e.g., Huang et al., 2010; Posid and Cordes, 2018; Wagner et al., 2019; O'Rear et al., 2020). The nuances in number knowledge that the Point-to-X task captures may allow researchers to understand the mechanism for acquiring number words. For example, future work could use Point-to-X to predict how soon children advance from one knower-level to the next.

How Do Children Acquire Number Words?

Questions about how children acquire the meanings of number words and the mechanisms for such a feat are core to the field of math cognition. Some accounts suggest that the ANS provides the basis for this process, where number words are mapped onto the imprecise representations of those quantities, with mapping progressing toward refinement with age (e.g., Gallistel and Gelman, 2000; Dehaene, 2009; Sasanguie et al., 2013; Starr et al., 2013; Odic et al., 2015). Others suggest that this process occurs through parallel individuation of objects and bootstrapping of prior number knowledge (e.g., Le Corre and Carey, 2007; Gunderson et al., 2015; Carey et al., 2017).

Our findings suggest that toddlers have some understanding of number words prior to learning their precise meanings. Although better able to map number words to small quantities, they nonetheless perform significantly above chance for all trial types queried here. However, the lack of distance effects in our results suggests that the mechanism for discriminating

quantities and mapping the number words here does not rely solely on the ANS. Barner (2012, 2017) suggests that the process of learning numbers words may entail two separate problems: First, children must learn to map number words to small numbers using cues, like linguistic number markings (singular/plural) and syntactic bootstrapping (Bloom and Wynn, 1997), and then eventually learn to associate large number words in their count list with approximate magnitudes.

Most previous work on mechanisms for acquiring number words has focused on explaining how children transition from being subset-knowers to cardinal principle knowers. This work typically focuses on older children who have acquired knowledge of multiple numbers, with less attention to toddlers at the cusp of understanding number words. Our findings suggest that toddlers have some preliminary understanding of number words above their knower-level, but this may only apply to children who have moved beyond knowing a single number (i.e., 2-knowers+).

Limitations, Conclusions, and Future Directions

Certain limitations warrant discussion. A large number of children did not complete the task due to inattention or outright refusal, which is common when testing infants and toddlers generally (e.g., Wynn, 1992; see Slaughter and Suddendorf, 2007 for review of this issue in infancy) but leaves unknown whether those children may show different patterns of number knowledge and Point-to-X performance than children included in analyses. Although Point-to-X may validly assess toddlers' number knowledge, other methods (such as looking-time) might reduce task demands and make the task more accessible to young children. Finally, our remote assessments of Point-to-X relied on physical materials being sent to the families' homes. We made this decision because families received physical materials for the Give-N task anyway and adding the Point-to-X materials did not result in any additional costs. By asking children to point to pages in front of them rather than images on the screen, parents could angle their webcams so that the researcher could see more easily what children pointed to. It is an open question whether a complete remote administration where children point to images on a screen shared by the researcher would work equally well.

Nonetheless, toddlers are able to successfully map number words to their referred quantities, even without fully understanding those number words. The Point-to-X task proves to be a flexible method for measuring children's number knowledge in-person and remotely, capturing nuances in children's number knowledge, and elucidating the mechanisms by which children acquire number word meanings. Future work using this task, especially using remote testing to reach families not typically represented in developmental research, might advance our understanding of children's early number knowledge and the acquisition of the cardinal principle.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be available at: <https://osf.io/ucyjg/>

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Pittsburgh Institutional Review Board, New York University Institutional Review Board, and University of Maryland Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

All authors contributed to conception and design of the study, manuscript revision, read, and approved the submitted version. AS performed the statistical analysis and wrote the first draft of the manuscript.

REFERENCES

- Agarwal, S., and Mazumder, B. (2013). Cognitive abilities and household financial decision making. *Am. Econ. J. Appl. Econ.* 5, 193–207. doi: 10.1257/app.5.1.193
- Almoammer, A., Sullivan, J., Donlan, C., Marušič, F., O'Donnell, T., and Barner, D. (2013). Grammatical morphology as a source of early number word meanings. *Proc. Natl. Acad. Sci. U. S. A.* 110, 18448–18453. doi: 10.1073/pnas.1313652110
- Barner, D. (2012). Bootstrapping numeral meanings and the origin of exactness. *Lang. Learn. Dev.* 8, 177–185. doi: 10.1080/15475441.2012.635541
- Barner, D. (2017). Language, procedures, and the non-perceptual origin of number word meanings. *J. Child Lang.* 44, 553–590. doi: 10.1017/S0305000917000058
- Bloom, P., and Wynn, K. (1997). Linguistic cues in the acquisition of number words. *J. Child Lang.* 24, 511–533. doi: 10.1017/S0305000997003188
- Carey, S. (2009). Where our number concepts come from. *J. Philos.* 106, 220–254. doi: 10.5840/jphil2009106418
- Carey, S., Shusterman, A., Haward, P., and Distefano, R. (2017). Do analog number representations underlie the meanings of young children's verbal numerals? *Cognition* 168, 243–255. doi: 10.1016/j.cognition.2017.06.022
- Cordes, S., and Gelman, R. (2005). "The young numerical mind: when does it count?" in *Handbook of Mathematical Cognition*. ed. I. D. Campbell (Psychology Press), 127–142.
- Currie, J., and Thomas, D. (2001). Early test scores, school quality and SES: longrun effects on wage and employment outcomes. *Res. Labor Econ.* 20, 103–132. doi: 10.1016/S0147-9121(01)20039-9
- Dehaene, S. (2009). Origins of mathematical intuitions: the case of arithmetic. *Ann. N. Y. Acad. Sci.* 1156, 232–259. doi: 10.1111/j.1749-6632.2009.04469.x
- Dehaene, S., Dehaene-Lambertz, G., and Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends Neurosci.* 21, 355–361. doi: 10.1016/S0166-2236(98)01263-6
- Dehaene, S., and Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition* 43, 1–29. doi: 10.1016/0010-0277(92)90030-L
- Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2007). School readiness and later achievement. *Dev. Psychol.* 43, 1428–1446. doi: 10.1037/0012-1649.43.6.1428
- Elliott, L., Braham, E. J., and Libertus, M. E. (2017). Understanding sources of individual variability in parents' number talk with young children. *J. Exp. Child Psychol.* 159, 1–15. doi: 10.1016/j.jecp.2017.01.011

FUNDING

This work was funded by the National Science Foundation (DRL1920545 to ML, HB, and EV-D, HRD1760844 to ML, HRD1760643 to NC, and HRD1761053 to CT-L). AS was supported by the National Institutes of Health under grant no. T32GM081760, and ML was supported by a Scholar Award from the James S. McDonnell Foundation.

ACKNOWLEDGMENTS

We thank Juliana Kammerzell, Danielle Fox, Erica Schweitzer, Margaret Isaacson, Olivia Sidoti, Taylor Montue, Natalie Heywood, Margaret Laird, Linsah Coulanges, Shirley Duong, Jessica Ferraro, Darcy Smith, Daniel Suh, Alexandra Mendelsohn, Yu (Tina) Chen, Valerie Mejia, Heidi Fuentes, and the research assistants in the Kids' Thinking Lab for the help with data collection and data entry. Finally, we especially thank the families who participated.

- Feigenson, L., and Carey, S. (2003). Tracking individuals *via* object-files: evidence from infants' manual search. *Dev. Sci.* 6, 568–584. doi: 10.1111/1467-7687.00313
- Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. *Trends Cogn. Sci.* 8, 307–314. doi: 10.1016/j.tics.2004.05.002
- Frye, D., Braisby, N., Lowe, J., Maroudas, C., and Nicholls, J. (1989). Young children's understanding of counting and cardinality. *Child Dev.* 60, 1158–1171. doi: 10.2307/1130790
- Gallistel, C. R., and Gelman, R. (2000). Non-verbal numerical cognition: from reals to integers. *Trends Cogn. Sci.* 4, 59–65. doi: 10.1016/S1364-6613(99)01424-2
- Gunderson, E. A., and Levine, S. C. (2011). Some types of parent number talk count more than others: relations between parents' input and children's cardinal-number knowledge. *Dev. Sci.* 14, 1021–1032. doi: 10.1111/j.1467-7687.2011.01050.x
- Gunderson, E. A., Spaepen, E., and Levine, S. C. (2015). Approximate number word knowledge before the cardinal principle. *J. Exp. Child Psychol.* 130, 35–55. doi: 10.1016/j.jecp.2014.09.008
- Halberda, J., and Feigenson, L. (2008). Development change in the acuity of the "number sense": the approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Dev. Psychol.* 44:1457. doi: 10.1037/a0012682
- Huang, Y. T., Spelke, E., and Snedeker, J. (2010). When is four far more than three? Children's generalization of newly acquired number words. *Psychol. Sci.* 21, 600–606. doi: 10.1177/0956797610363552
- Ip, M. H. K., Imuta, K., and Slaughter, V. (2018). Which button will I press? Preference for correctly ordered counting sequences in 18-month-olds. *Dev. Psychol.* 54, 1199–1207. doi: 10.1037/dev0000515
- Jordan, N. C., Kaplan, D., Ola, L. N., and Locuniak, M. N. (2006). Number sense growth in kindergarten: a longitudinal investigation of children at risk for mathematics difficulties. *Child Dev.* 77, 153–175. doi: 10.1111/j.1467-8624.2006.00862.x
- Le Corre, M., and Carey, S. (2007). One, two, three, four, nothing more: an investigation of the conceptual sources of the verbal counting principles. *Cognition* 105, 395–438. doi: 10.1016/j.cognition.2006.10.005
- Le Corre, M., Van de Walle, G., Brannon, E. M., and Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cogn. Psychol.* 52, 130–169. doi: 10.1016/j.cogpsych.2005.07.002
- Levine, S. C., Suriyakham, L. W., Rowe, M. L., Huttenlocher, J., and Gunderson, E. A. (2010). What counts in the development of young children's number knowledge? *Dev. Psychol.* 46, 1309–1319. doi: 10.1037/a0019671

- Libertus, M. E., and Brannon, E. (2009). Behavioral and neural basis of number sense in infancy. *Curr. Dir. Psychol. Sci.* 18, 1457–1465. doi: 10.1111/j.1467-8721.2009.01665.x.Behavioral
- Navarro, M. G., Braham, E. J., and Libertus, M. E. (2018). Intergenerational associations of the approximate number system in toddlers and their parents. *Br. J. Dev. Psychol.* 36, 521–539. doi: 10.1111/bjdp.12234
- Odic, D., Le Corre, M., and Halberda, J. (2015). Children's mappings between number words and the approximate number system. *Cognition* 138, 102–121. doi: 10.1016/j.cognition.2015.01.008
- O'Rear, C. D., McNeil, N. M., and Kirkland, P. K. (2020). Partial knowledge in the development of number word understanding. *Dev. Sci.* 23:e12944. doi: 10.1111/desc.12944
- Posid, T., and Cordes, S. (2018). How high can you count? Probing the limits of children's counting. *Dev. Psychol.* 54, 875–889. doi: 10.1037/dev0000469
- Reyna, V. F., and Brainerd, C. J. (2007). The importance of mathematics in health and human judgment: numeracy, risk communication, and medical decision making. *Learn. Individ. Differ.* 17, 147–159. doi: 10.1016/j.lindif.2007.03.010
- Sasanguie, D., Göbel, S. M., Moll, K., Smets, K., and Reynvoet, B. (2013). Approximate number sense, symbolic number processing, or number-space mappings: what underlies mathematics achievement? *J. Exp. Child Psychol.* 114, 418–431. doi: 10.1016/j.jecp.2012.10.012
- Slaughter, V., and Suddendorf, T. (2007). Participant loss due to “fussiness” in infant visual paradigms: a review of the last 20 years. *Infant Behav. Dev.* 30, 505–514. doi: 10.1016/j.infbeh.2006.12.006
- Starr, A., Libertus, M. E., and Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proc. Natl. Acad. Sci. U. S. A.* 110, 18116–18120. doi: 10.1073/pnas.1302751110
- StataCorp (2017). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.
- Trusty, J., Robinson, C. R., Plata, M., and Ng, K.-M. (2000). Effects of gender, socioeconomic status, and early academic performance on postsecondary educational choice. *J. Couns. Dev.* 78, 463–472. doi: 10.1002/j.1556-6676.2000.tb01930.x
- van Marle, K., Chu, F. W., Li, Y., and Geary, D. C. (2014). Acuity of the approximate number system and preschoolers' quantitative development. *Dev. Sci.* 17, 492–505. doi: 10.1111/desc.12143
- Wagner, K., Chu, J., and Barner, D. (2019). Do children's number words begin noisy? *Dev. Sci.* 22:e12752. doi: 10.1111/desc.12752
- Wang, J., and Feigenson, L. (2019). Infants recognize counting as numerically relevant. *Dev. Sci.* 22:e12805. doi: 10.1111/desc.12805
- Wynn, K. (1990). Children's understanding of counting. *Cognition* 36, 155–193. doi: 10.1016/0010-0277(90)90003-3
- Wynn, K. (1992). Children's acquisition of the number words and the counting system. *Cogn. Psychol.* 24, 220–251. doi: 10.1016/0010-0285(92)90008-P

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Silver, Elliott, Braham, Bachman, Votruba-Drzal, Tamis-LeMonda, Cabrera and Libertus. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

Point-to-X Task Stimuli

Word-control practice trials.

| Prompt | Image 1 | Image 2 |
|---------------------|---------|---------|
| "Which has a tree?" | Tree | Cup |
| "Which has a ball?" | Banana | Ball |

Number-word trials.

| Prompt | Image 1 | Image 2 |
|-----------------------------|---------|---------|
| "Which has 1 cookie?" | 1 | 3 |
| "Which has 2 fish?" | 7 | 2 |
| "Which has 4 ducks?" | 4 | 5 |
| "Which has 5 apples?" | 5 | 1 |
| "Which has 2 carrots?" | 2 | 4 |
| "Which has 3 ladybugs?" | 10 | 3 |
| "Which has 4 strawberries?" | 3 | 4 |
| "Which has 5 pears?" | 5 | 3 |
| "Which has 10 fish?" | 4 | 10 |
| "Which has 3 oranges?" | 2 | 3 |
| "Which has 7 blueberries?" | 7 | 5 |
| "Which has 1 turtle?" | 2 | 1 |



Studying Children's Eating at Home: Using Synchronous Videoconference Sessions to Adapt to COVID-19 and Beyond

Shruthi Venkatesh* and Jasmine M. DeJesus

Department of Psychology, University of North Carolina at Greensboro, Greensboro, NC, United States

OPEN ACCESS

Edited by:

Rhodri Cusack,
Trinity College Institute of
Neuroscience, Ireland

Reviewed by:

Rajagopal Raghunathan,
University of Texas at Austin,
United States
Eleonora Ceccaldi,
University of Genoa, Italy
Shelley Van Der Veeke,
Leiden University, Netherlands

*Correspondence:

Shruthi Venkatesh
s_venkat@uncg.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 April 2021

Accepted: 30 June 2021

Published: 22 July 2021

Citation:

Venkatesh S and DeJesus JM (2021)
Studying Children's Eating at Home:
Using Synchronous Videoconference
Sessions to Adapt to COVID-19
and Beyond.
Front. Psychol. 12:703373.
doi: 10.3389/fpsyg.2021.703373

The COVID-19 pandemic has disrupted many facets of developmental research, including research that measures children's eating behavior. Here, children's food intake is often measured by weighing foods that children are offered before and after in-person testing sessions. Many studies also examine children's food ratings (the extent to which they like or dislike a food), assessed *via* picture categorization tasks or hedonic scales. This paper reviews existing research on different methods for characterizing children's eating behavior (with a focus on food intake, preferences, and concepts) and presents a feasibility study that examined whether children's eating behaviors at home (including their food intake and ratings) can be measured *via* live video-chat sessions. The feasibility analyses revealed that an observational feeding paradigm at home yielded a majority (more than 70%) of video-chat recordings that had a sufficient view of the child and adequate sound and picture quality required for observational coding for the majority of the session's duration. Such positioning would enable behavioral coding of child food intake, parent food talk, and meal characteristics. Moreover, children were able to answer questions to stories and express their preferences *via* researcher screen-share methods (which can assess children's self-reported food preferences and beliefs) with low rates of exclusion across studies. The article ends with a discussion on the opportunities and challenges of using online platforms to conduct studies on children's eating behaviors in their home environments during the COVID-19 pandemic and beyond.

Keywords: meal observation, children's eating behavior, online research, food preferences, food intake

STUDYING CHILDREN'S EATING AT HOME: ADAPTING TO COVID-19 AND BEYOND

COVID-19 has upended many aspects of the research process (not to mention the lives of researchers and the families we study). Before the pandemic, researchers ascertained the validity of remotely collecting data from children of a variety of ages using asynchronous measures, including webcam recorders for looking-time paradigms with infants (Simmelmman et al., 2017; Tran et al., 2017) and unmoderated research platforms, such as LookIt (Scott and Schulz, 2017) and Discoveries in Action (Rhodes et al., 2020). These platforms allow children to

complete studies without interacting with a researcher directly (but with some assistance from the parent or guardian providing consent). Many of these methods have been recommended during the pandemic to continue and potentially improve data collection into the future (Sheskin et al., 2020). Synchronous methods of remote data collection, such as Zoom, have also become popular as they allow researchers to interact with and collect data from families in live time (Kuo et al., 2021). However, there is limited work on the feasibility of studying children's eating behavior (a key line of research in our laboratory) using remote online research tools. In this paper, we document the successes and challenges we have experienced in adapting our research using online methods. In the upcoming sections, we highlight previous work that has measured infants' and children's eating behavior using the amount of food eaten and food preferences or concepts as outcome measures in the laboratory or outside the laboratory in home or school settings. We present data from a feasibility study that examined children's typical meals at home and food preferences *via* live video-chat sessions. We conclude with a discussion on the opportunities to ask innovative questions about children's eating behavior at home during and after the COVID-19 pandemic using such online platforms.

Measuring Children's Food Intake

Comparing Pre- and Post-test Food Weight

Prior to the COVID-19 pandemic, researchers interested in examining children's food intake took a variety of approaches in measuring what and how much children ate during a research study. A common and intuitive approach to this practice was to weigh a food that children were offered before the study and weigh that food again after the study as a measure of how much children ate. In a comprehensive review on experimental studies that seek to change children's eating behavior, 29 of the 120 studies reviewed used weighed food intake as a dependent variable (among other common outcomes, such as food preferences or choices which will be described in the upcoming sections), specifically for studies that sought to increase children's fruit and vegetable intake (see DeCosta et al., 2017 for review). Many of our own studies take this approach, including studies that examine how social knowledge of the food influences children's food intake (DeJesus et al., 2018b), whether children eat more food if they assisted in preparing the food (DeJesus et al., 2019a), how maternal talk and intake of food relates to children's intake of those foods (DeJesus et al., 2018a), and whether children learn about food by verbal testimony or by seeing someone eat that food (DeJesus and Venkatesh, 2020). When this in-person interaction is not possible, it is harder for researchers to use pre-post weight measurements as a standardized measure of food intake given the access to and variability of weighing scales that families may have at home.

Measuring Food Intake *via* Bites or Pieces of Food Eaten

In addition to measuring intake based on food weight, researchers can code the number of bites (solid intake) of food taken

during feeding sessions which can be coded from video recordings. For older infants and children who eat solid foods, food bites as an outcome variable are indexed by coding for every time the food passes through the children's lips. As an example, to validate maternal reports of their child's selective eating against children's observed food intake, Fernandez et al. (2018) examined data from an observational paradigm during which familiar and unfamiliar foods were offered. Researchers measured the children's latency to their first bite of food and the total number of bites in the videos by counting the number of times the food passed through the infant's lips in 10-s intervals (Fernandez et al., 2018). Similarly, in a study examining one-year old infants' temperament and feeding history as predictors of their receptivity to unfamiliar foods, infant's acceptance of the food was coded from videos in 5-s intervals (Moding et al., 2014). Here, acceptance was defined by when the infants opened their mouths in anticipation of the next bite, smiled and reached toward the food, or the food successfully passed through their lips. Food rejection was coded when the infants physically removed the foods from their mouths, fussed, or turned their mouths away. Intake in bites can also be captured in terms of children's choice of one food over another (e.g., do children take their first bite of food A or food B?), where the foods that the infants reach toward and taste first are measured (Shutts et al., 2009). Thus, food bites can be one avenue through which researchers can gather quantitative information on food intake, and we aimed at exploring whether such data can be collected through online data collection methods.

Another quantitative measure of food intake is counting the number of discrete pieces of food eaten. For example, if a child is offered 10 carrot sticks, how many carrot sticks did the child eat? In an intervention that sought to conceptually explain food as a source of nutrition to preschool children, researchers live coded children's snack intake during snack time at their preschool setting (Gripshover and Markman, 2013). The authors found an increased intake of vegetables post the intervention in children; here, the number of pieces of snack consumed was measured by number of pieces of food chosen minus those left after the snack time (such as crackers and vegetables). Comparably, to test the IKEA effect, or the idea that people prefer self-crafted products over similar products made by others (Norton et al., 2012), in children, Raghoobar et al. (2017) explored whether children would consume more vegetables if they created the snack themselves. Children crafted a peacock out of either snack vegetables or colored beads and their vegetable intake was measured by the number of vegetable pieces (e.g., cucumber) pre-post intake.

An extension of this method to examine food choices is to assess children's choices when the same food is presented in different conditions. To investigate whether the knowledge that a food is healthy or can help with an intellectual goal will imply that the food tastes less good, 3- to 5.5-year-old children were offered either crackers or carrots across five experiments. Based on their condition, they received "healthy," "yummy," and control (no message) messages of the food, with the amount eaten (in terms of pieces), the number of pieces

of food chosen to take home, and perceived ratings of the food as the dependent variables (Maimaran and Fishbach, 2014). Such coding eliminates the added personnel power, software, and time needed for coding bites as described previously and can be completed live during the testing session. However, the number of foods that can be counted as discrete pieces is restricted in comparison with the variety of textures and forms of food infants and children consume in their home environments.

Measuring Food Preferences and Concepts

In addition to actual food intake, children's food preferences can also be assessed, either in addition to their food intake or as a primary outcome without offering children real foods. Such studies typically highlight children's understanding of food groups, their own food preferences, and their other beliefs about food, such as potential connections between food and cultural groups. Children's verbal attestation of their food preferences, likes and dislikes can be measured through picture choices, brief scale ratings, and sorting tasks. Children can be asked to report their preference on a scale through smiley face rating scales (ranging from "not yummy at all" to "really really yummy" or "dislike" to "like;" Zeinstra et al., 2010; DeJesus et al., 2018b), a series of questions, such as "Is [name of food] yummy or yucky? Really (yummy/yucky) or a little (yummy/yucky)?" (DeJesus et al., 2019b), or as a choice between two options (Echelbarger et al., 2020). For preverbal infants who cannot say if they like a food or not, a few methods are still available to assess their preferences or early reasoning about food: infants' facial expressions or parent ratings can provide some insight into their food enjoyment (Mennella et al., 2001).

Similar methods can be used to understand children's thinking about other aspects of food, such as their social relevance and taxonomic categories. For instance, when presented with pictures of foods that included conventional and unconventional combinations, in addition to their own preferences, children's social judgments about people who ate those foods were assessed with questions, such as "do you want to be friends with [name of person who eats that food] or not really?" (DeJesus et al., 2019b). Social judgments can even be assessed in preverbal infants using looking-time paradigms, such as examining whether people who share a food preference are especially likely to socially affiliate (Lieberman et al., 2016). Finally, card sorting tasks have been used to examine children's ability of food categorization as a precursor to food rejection (Rioux et al., 2016). Here, children were shown pictures of fruits and vegetables that varied in color, typicality, and whether the foods had been cubed or sliced. Children completed tasks, such as sorting those pictures into categories, naming the colors of the fruits and vegetables, and discarding foods they were unwilling to try (Rioux et al., 2016). In these ways, researchers can assess infants and children's food preferences and ratings verbally and nonverbally, through picture choices, brief scale ratings, sorting tasks, and looking time paradigms. In this paper, we hoped to examine

the feasibility of collecting children's self-reported preferences *via* an online format.

Parental Reports of Children's Food Intake

Parental recall and reports of their children's diet can provide descriptive data on what kinds of food their children eat, which can be standalone data and predictors or outcomes in studies that also have meal observations. In a study that combined naturalistic home meal recordings with parental report data, parents reported on their toddlers' food intake *via* three 24-h dietary recall interviews, and the foods stated were later coded into food groups, specifically fruits and vegetables (Edelson et al., 2016). Videos of meals at home over a day were collected and children's acceptance or refusal of the foods were coded, along with parental food talk language (prompts). Among other findings, the more fruit and vegetable prompts parents used during the recorded meals, the more parents reported their child ate fruits and vegetables in a 24-h dietary recall task (Edelson et al., 2016). While parent recall was used as an outcome variable in this study, such reports can also be used as predictor variables. Indeed, in a longitudinal study examining infant growth trajectories, mothers reported on their infants' food frequency and milk (breast milk, formula, or other milks) intake across 7 days as a predictor of child obesity at 6 years (measured by BMI) with infants' change in weight-for-length z-scores over the first year post-partum as the mediator of this relationship (Ventura et al., 2020).

In addition to parent dietary recall, parent reports on their children's eating habits and dietary patterns are another common source of data. As an example, the Child Eating Behavior Questionnaire developed by Wardle et al. (2001) consists of subscales, such as food responsiveness, children's food fussiness, children's emotional over and undereating, food enjoyment, desire to drink, and satiety responses. This scale of parent report that can be used to predict children's obesity has been validated against behavioral measures of children's obesogenic behaviors (Carnell and Wardle, 2007). Furthermore, the Comprehensive Feeding Practices Questionnaire is another commonly used parent-report measure that contains 12 subscales of feeding practices, such as using food as a reward, routine of eating, and teaching nutrition (Musher-Eizenman and Holub, 2007). This questionnaire can be administered *via* paper-pencil or online survey, which lends its flexibility for being used in different settings. In these ways, parents can not only provide data on their children's eating behaviors, but can also help in collecting such data *via* online formats, which will be elucidated in our Methods and Discussion.

The Present Study: Feasibility of Measuring Food Intake and Ratings Online

Prior research provides multiple methods to study children's eating behavior, including naturalistic video recordings that capture children's eating at home. However, there is a dearth of research that analyzes the validity and plausibility of adapting

TABLE 1 | Child racial and ethnic distribution (Method 1–2).

| | Method 1 | Method 2 |
|-----------------------------------|------------------|-------------------|
| | (<i>n</i> = 50) | (<i>n</i> = 181) |
| Latinx | 2 (4%) | 13 (7%) |
| Caucasian/White | 40 (80%) | 104 (57%) |
| African-American | 1 (2%) | 10 (6%) |
| Asian/Asian-American | 1 (2%) | 32 (18%) |
| More than one race | 5 (10%) | 9 (5%) |
| Prefer not to respond/no response | 1 (2%) | 13 (7%) |

TABLE 2 | Primary parent education (Method 1).

| | Frequency |
|------------------------------|-----------|
| High school/GED | 1 (2%) |
| Associate's degree | 3 (6%) |
| Bachelor's degree | 7 (14%) |
| Some graduate school | 2 (4%) |
| Graduate/professional degree | 36 (72%) |
| Other | 1 (2%) |

these measures to remotely study children's eating behavior using synchronous videoconference sessions. The COVID-19 pandemic has disrupted our ability to invite families into the laboratory for a feeding experimental study or even manage the personnel required for home video recordings. With the shift of our field toward remote online data collection over the course of this past year, our laboratory also transitioned to collecting data from children and families through synchronous videoconference sessions as we describe in two methods. In Method 1, we describe the online remote methods to observe children's typical meal times at home, and the likelihood of being able to code certain behaviors from these video recordings, such as whether coders could see the children's face and mouth and hear the parent's talk during the session. In Method 2, we describe a synchronous videoconference method that could be used to attain children's food ratings, categorizations, or other aspects of their reasoning about food.

METHOD 1

Observations of Eating at Home

Video recordings of young children's meals at home have yielded information about the characteristics of the family meal and parental food talk (Bergmeier et al., 2015b). Moreover, videos have also been a method through which their actual food intake has been coded, by measuring liquid sucking, food bites, and behaviors related to acceptance or rejection of foods (Lumeng et al., 2007; Moding et al., 2014; Fernandez et al., 2018). The goal of this current study was twofold. First, we aimed at collecting pilot data to examine typical meals at home for children under 3 years of age and assess the feasibility of conducting these studies using synchronous videoconference sessions. Second, we hoped to test the plausibility of conducting

an experimental manipulation of feeding behaviors in an environment naturalistic to the child, which could be an externally valid approach even beyond the COVID-19 pandemic.

Participants

Children under the age of 3 years were recruited for this study. Participants were recruited from an existing database of volunteer families, social media advertising, and Children Helping Science, an online platform for researchers to advertise online studies and for parents to sign up for studies. Parents were informed *via* email that we would like to set up a half-hour videoconference during their child's typical meal or snack time, and the appointment was scheduled accordingly (parents were given the flexibility to choose what meal was observed). We were predominantly interested in testing infants as they transitioned to solid foods and toddlers as they expanded their repertoire of solid foods, which is why this age range was chosen. We also aimed to offer an activity for younger siblings of children participating in other research projects designed for children aged 3 years and older.

We had 50 children (25 females, $M_{\text{age}} = 17.88$ months, $\text{Range}_{\text{age}} = 0\text{--}55$ months) participate in the study, with three sibling pairs who participated in the same session together. Though the target age for this study was 3 years and under, one child in the sibling pair was 4 years old and was eating a meal along with their younger sibling. Since this was a typical setup for the family, the older child's data were retained. Parents identified the majority of our child sample as not Hispanic/Latino (47 or 94%) and as Caucasian/White (42 or 84%; see **Table 1**). Parental demographics indicate that 36 (72%) parents had graduate degrees, and 26 (52%) reported combined annual household income to be more than \$120,000 (see **Tables 2 and 3**). All parents reported English as a language spoken at home, and 14 (28%) reported a secondary language (such as Russian or French). Since this is a feasibility study, we sought to retain all participant videos to document the range and frequency of issues that would potentially hinder behavioral coding. However, we had decided to exclude videos if they were so poor in quality that even the feasibility analysis (described under "Descriptive Data of the Feeding Sessions") could not be extracted from these videos. Our other exclusionary criteria included if children were distressed by the presence of the video recording device. None of the sessions fit these criteria, hence, we did not exclude any video recordings.

Materials and Procedure

Once the videoconference appointment was scheduled, parents were emailed an online consent form. This form also included a media consent form which gave us permission to videotape this interaction and potentially use the audio and video recordings (such as at conferences, for teaching materials, or on our laboratory Web site). All parents consented to being recorded, though there was variability in the permissions granted for the use of these recordings (see **Table 4**). Parents reported on demographics, such as their race, ethnicity, educational attainment, household income, and languages spoken at home.

TABLE 3 | Combined household income (Method 1).

| | Frequency |
|-----------------------|-----------|
| Less than \$15,000 | 2 (4%) |
| \$25,000–\$40,000 | 2 (4%) |
| \$40,000–\$60,000 | 6 (12%) |
| \$60,000–\$90,000 | 5 (10%) |
| \$90,000–\$120,000 | 5 (10%) |
| More than \$120,000 | 26 (52%) |
| Prefer not to respond | 4 (8%) |

TABLE 4 | Parent media permission (Method 1).

| | Yes | No | Missing |
|---|----------|----------|---------|
| Showing videos, audio, or images in the classroom | 42 (84%) | 7 (14%) | 1 (2%) |
| Showing videos, audio, or images in academic meetings or conferences | 40 (80%) | 9 (18%) | 1 (2%) |
| Showing videos, audio, or images on our laboratory Web site | 24 (48%) | 25 (50%) | 1 (2%) |
| Including images in publications of this study and on online repositories, such as the Open Science Framework | 27 (54%) | 22 (44%) | 1 (2%) |
| Including images in newsletter we send to families interested in our research | 28 (56%) | 21 (42%) | 1 (2%) |
| Including images in promotional materials (such as brochures or flyers) | 27 (54%) | 22 (44%) | 1 (2%) |

In this email, parents were also sent a guide to help navigate them through the video-chat platform if needed (full text available on the Open Science Framework).¹ This guide contained screenshots for how participants could join the meeting and turn on their video and audio settings. We used WebEx when our online data collection began in May 2020 as our institution's IRB had already approved research studies using that platform. By October, we learned that our university would be ending its subscription with WebEx and we transitioned to using Zoom for data collection. Zoom was also more familiar to parents (a few parents asked if we could use Zoom instead) and was an easier platform to use (though we did not experience any technical difficulties that resulted in participant exclusion specifically because of difficulties with the WebEx platform).

Researchers conducted the study on a university-issued laptop or desktop device. Parents typically logged in from their laptops,

but they also could log in from their tablet, phone, or desktop computer. At the start of videoconference session, the researcher introduced the study to the parent and started recording the session. The recording was done directly to the device the researcher was logged in on and not on the WebEx/Zoom cloud for participant privacy. Parents were asked some questions before the start of the feeding session regarding what their child was going to eat, if the child would be sitting in their typical seat, how often they had been introducing new foods to their child during the pandemic, and if there was anything about the current pandemic situation they would like to share (*see OSF for full text*). Once they were ready to start with the feeding session, the researcher suggested that parents could cover their screen with a sheet of paper (without covering the camera), or swipe to another application on their device if the child seemed distracted by seeing themselves eat or if eating in front of a screen was atypical for them. If the parents chose to do this, it was ensured that the camera view of the feeding setting was not blocked. The researcher then told the parents to “do what you would usually do as if we were not there” and told the parent they would return if the parent said they were done with the session or after 30 min had passed. The researcher did not provide any additional setup instructions to the parents, as the goal was to assess the quality of the videos that could be recorded with minimal researcher guidance. The researcher then muted/turned off their video and started a 30-min timer.

During the videoconference session, the researcher made live notes of some characteristics of the feeding session, such as if the parent–child dyad was in the frame, if the food was visible, whether it was an individual or family meal, and whether the child was self-feeding or being fed (*see OSF for full text*). After 30 min passed or the researcher heard the parent say, “we are done” (whichever came first), the researcher then turned their video back on and unmuted, and to conclude, asked the parent whether they noticed any differences from a typical meal and if there was anything else they thought would be important for us to know. The child was emailed a certificate and an age-appropriate e-story book from the “Amazing Books for Children” series by the Center for Disease Control and Prevention.² The recorded video was then uploaded to our laboratory's secure Box folder. This study and the study described under Method 2 were conducted in 2020–21 and approved by the University's Institutional Review Board (20–0365, “Online child development studies”). Deidentified data and relevant research materials are available on the Open Science Framework (see footnote 1).

Video Issues Coding

The goal of this study was to document the feasibility of assessing children's feeding behaviors *via* recordings of synchronous video-chat sessions. Specifically, we intended to illustrate the plausibility of coding child food intake in bites and parental speech and behavior during meals. To this end,

¹<https://osf.io/rhmuq/>

²<https://www.cdc.gov/ncbddd/actearly/amazingme.html>

we developed a coding scheme to record potential issues in these video recordings, or a characteristic in the recorded feeding time that would interfere with our behavioral coding goals. We identified 10 types of issues that could appear in these video recordings: (1) cannot see parent's face (2) cannot see child's face (3) parent's hand comes in front of child's mouth (4) video too dark (5) audio not clear (6) child's mouth blocked during bottle feeding (7) cannot see individual children (when more than one child participated at once) (8) child moves in and out of frame (9) some speech not in English, and (10) Internet connectivity issues (*see OSF for full text*).

First, we stated if each of these issues was present in the video or not. If it was present, then we quantified the severity of the issue, or for how long in the feeding session the issue occurred. For example, if a researcher intended to code maternal engagement with the child during the feeding session, and the mother was in the frame for most of the video but stepped out of the frame for a few minutes to refill the child's plate, the coding would still be possible for most of the session. In contrast, if the child was sitting in front of a window and was backlit for the whole meal, then it would be harder to code their food intake or parent-child engagement.

For each issue, we coded whether it occurred for the whole video (100%), most of the video (75%), about half the video (50%), little of the video (25%), or not at all (i.e., it was not an issue in the video). These degrees of severity were estimated based on the duration of the feeding session. For instance, if a feeding session was about 20-min long, we noted first if the issue occurred or not. If it did occur, then we saw whether it occurred for little of the video (5 min), half the video (10 min), most of the video (15 min), or the whole time (20 min). For brightness of the video, we added an additional code "can still see child and food set-up, but brightness is not great" as a comparison for videos that were very clear in terms of visibility to those that were less clear. For bottle feeding and parent language, we coded the presence of these issues given the proportion of time that the behavior occurred. For instance, the mother could be talking for the whole duration to other family members in addition to the child. We coded the language the mother talked to the child in and, if bilingual, assessed the proportion of time the mother did not speak in English to the child. Similarly, if children had bottle feeds (milk/water) during their solid food sessions, for example, they drank out of a bottle for 3 min of a 15-min meal session, then we coded whether their mouth was blocked or not during those 3 min. A team of four coders established inter-rater reliability for 20% of the dataset and had inter-class Kappas of at least 0.76 for each code.

Results

Descriptive Data of the Feeding Sessions

One parent participated when their infant was bottle fed at 3 months, and again 4 months later when the infant had transitioned to solid foods. For the analysis to follow, we included both their videos as a measure of bottle and solid feeds. Seven

parents scheduled a session and filled out the consent form but did not attend or reschedule the appointment. Of the 48 videos ($n = 44$ individual child sessions, $n = 1$ child repeated at two time points, $n = 3$ sibling sessions), the mother attended the appointment for 41 sessions (85%), the father attended the appointment for three sessions (6%), and both parents attended the appointment for four sessions (8%). 33 sessions (69%) were individual meals where only the child was eating, while 15 (31%) were family meals, where we could see the child as well as other family members eating a meal. Furthermore, eight (16%) children were fed by the parent, 28 (55%) children self-fed, and 15 (30%) had a mix of both, self-feeding and being fed.

In terms of the type of feeding involved, three (6%) were only bottle feeds, while the majority (48 or 94%) was solid food sessions. 17 feeding sessions (35%) lasted the whole 30 min. From the sessions that did not last for 30 min (i.e., sessions that ended because the parent said they were done), 22 min was the average duration of the meal.

Parent Interview

With regards to the parents' description of the meal, all children sat in their typical seats during the meal. Since the start of the pandemic, 15 parents (30%) said they have been introducing new foods to their child more than usual, two (4%) less than usual, and 26 (51%) about the same pace as before. 27 (53%) parents described the session as representative of a typical meal. Some common responses for atypicality of the meal were "Normally my husband and I will talk to each other more during breakfast" or "we usually start with a food he [the baby] likes and then offer a new food, but we thought you would be interested in seeing him eat a new food so we started with that first."

Video Issues Coding

Child visibility. A majority of the videos did not contain issues that would potentially hinder behavioral coding (see Table 5). We could see the child's face for the whole session in 35 videos (73%), and in seven videos (15%), we could not see the child's face for only a little of the video (less than 25% of duration). For videos where parents fed their child, their hands did not cover the child's mouth at all in 44 sessions (92%). 40 children (83%) were seated in one place and did not move around (were in the video frame) for the entire video, and six (13%) moved around a little bit.

Parent visibility and language use. The data were mixed with regard to parents being in the frame. In 17 (35%) videos, the parents were in the frame the whole time, and in 15 (31%) videos, the parents were not in the frame at all. However, of the 20 videos in which the parent was not in the frame for most or all of the video, we could hear them talking in 18 (90%) videos. Parents spoke in English to their child in 44 videos (92%) and did not speak in English at all in two videos (4%).

TABLE 5 | Frequencies (%) of video issues in naturalistic videoconference meal time observations.

| | Not an issue | Little of the video | Half of the video | Most of the video | All of the video |
|---|--------------|---------------------|-------------------|-------------------|------------------|
| Cannot see parent's face | 17 (35) | 8 (17) | 3 (6) | 5 (10) | 15 (31) |
| Cannot see child's face | 35 (73) | 7 (15) | 2 (4) | 4 (8) | 0 |
| Parent's hand comes in front of child's mouth | 44 (92) | 1 (2) | 1 (2) | 2 (4) | 0 |
| Lighting issues (e.g., video too dark) | 37 (77) | 1 (2) | 0 | 0 | 0 |
| Audio not clear | 43 (90) | 4 (8) | 1 (2) | 0 | 0 |
| Child moving around | 40 (83) | 6 (13) | 1 (2) | 1 (2) | 0 |
| Bilingual/Not in English | 44 (92) | 0 | 1 (2) | 1 (2) | 2 (4) |
| Internet connectivity issues | 41 (85) | 7 (15) | 0 | 0 | 0 |

Percentages rounded to nearest whole number. For the lighting issues category, for 10 videos (21%), the video was coded as "brightness is not great but can still see food/child."

Food visibility. For 26 (51%) of the children, we could see their food directly, for 23 (45%) children, we could see their eating set up but not the food directly unless it was picked up, and in two (4%) sessions, the view of the food was obstructed. Of the 12 feeding sessions that included bottle feeds, the children's mouths were blocked by the bottle for most or all of the video in seven (58%) sessions.

General visibility and connectivity. In terms of visibility, 37 videos (77%) had good brightness for all of the video, followed by 10 videos (21%) where we could still see the feeding setup but the brightness was comparatively lower. The more challenging videos were the sibling studies when more than one child was eating together in the same session. Here, in all three of these sessions, we could not see individual children for most or all of the session which would interfere with food bites or individual eating behavior coding.

We also wanted to capture disruptions regarding to Internet connectivity. In 43 videos (90%), the audio was clear for the entire video, and in 41 videos (85%) there were no Internet connectivity issues. In seven videos (15%), Internet connectivity issues existed for a little (less than 25%) of the duration, which indicates brief freezing frames in the recording. In none of the videos was Internet connectivity an issue for the entire video (i.e., the family did not freeze completely, or we did not have to restart/cancel the session).

Discussion

This feasibility study revealed that, for the most part, observational meal recordings garnered through synchronous videoconference sessions yield codable data. Specifically, researchers can view the child's face, feeding setup, and food intake clearly, with reasonable audio and video quality and the child being seated in one place (i.e., not moving in and out of frame frequently). Although parents themselves were not present in these videos all the time, parent talk was recorded for subsequent coding (e.g., for researchers interested in parental prompts or other types of verbal engagement during meals). One potential reason why parents were not in the frame is because we did not explicitly tell them to be there. Parents interpreted our instructions differently, and hence, they were mixed in terms of who was visible in the frame (just the child or the parent and child), especially when the child self-fed. Similarly, we did not instruct parents as to what type of foods to feed their child, so some parents mentioned that they made their child's most liked food to ensure they have a smooth session with us, while others tried an unfamiliar food as they believed it would be interesting for us. However, whether or not this variability would count as an "issue" for researchers depends completely on their research questions and can be solved through live feedback from the researcher to the parent, a topic we return to in the General Discussion.

As observed, 15 sessions were family meals, where parents and other family members could be seen eating along with the child in these videos. The presence of family and companions facilitates greater food intake during mealtimes (De Castro, 1994). Moreover, seeing adults or peers socially modeling eating increases children's food intake (see Cruwys et al., 2015 for review). In this way, mealtime observations at home could offer the opportunity to study such social influences on food intake. However, this was not the focus of the present research as we aimed at assessing the feasibility of collecting data on children's individual eating. Moreover, we found that in the videos that had more than one child in the frame, the data quality was reduced as all children were not always in view. Here, the extent of the issue is also dependent on the device used by parents to call into the session and how far away the device was placed from the children. If parents call in from their tablet or mobile phone, then their camera view is narrower. If parents physically move the camera from one child to another to correct for this narrow view and attempt to capture both children, it is actually more difficult to extract any data (as each child is only visible for some of the session), compared to focusing on only one child (which means losing one child's data but having full data from another). If parents call in from a laptop device and place the laptop further away from the children to get a wider frame of the feeding setup, further distance reduces the ability to clearly view the food and child food bites. Hence, depending on one's research question, the presence (and absence) of other family members can be facilitated in such a setup that occurs at home.

Additionally, compared to solid food sessions, a majority of the bottle feeds obstructed the view of the child's mouth in the video recordings which would be challenging to code sucks. Therefore, it is critical for researchers to consider the type of data they hope to obtain, test out their videoconferences on multiple types of devices, and develop specific instructions to walk through with parents to capture the angles and information needed.

In addition to food bites, in Method 2 we describe the use of synchronous videoconference sessions to assess child food ratings and preferences.

METHOD 2

Asking Children Questions About Food

Apart from measuring actual food intake, another method of assessing children's eating behavior is eliciting their food ratings or beliefs about foods (e.g., DeJesus et al., 2019b; Echelbarger et al., 2020). In this section, we highlight online studies conducted during the COVID-19 pandemic that have assessed children's preferences and predictions as a plausible method for examining children's opinions about foods using synchronous videoconference sessions. Specifically, we briefly describe the methods, exclusion criteria, and attrition across studies.

Participants

We started data collection *via* synchronous videoconference sessions with children aged 3 to 12 years in May 2020 and have collected data from 192 children to date. We excluded 11 children's data (detailed under "Results"), which yielded a usable sample of 181 children (98 female). In addition, 10 parents scheduled an appointment but their child(ren) did not ultimately participate in the study (two parents completed the consent form but did not attend the appointment, seven parents did not complete the consent form nor attend the appointment, and one parent chose not to participate after learning that they would need to log in using video). Collapsing across studies, our participants identify as 13 (7%) Latinx and 104 (57%) Caucasian/White (see **Table 1**). Across studies, we follow a similar recruitment protocol to Method 1 (i.e., families are recruited *via* our volunteer database, social media advertising, and Children Helping Science). Parents were emailed the consent forms specific to their child's study in advance of the synchronous videoconference session. After the study, children were emailed a certificate and their choice of prize pack (an activity book of do-at-home science experiments, coloring sheets, word puzzles, recipes, or mazes) for participating.

Screen-Sharing Check Procedure

In these studies, researchers shared their screen with participants. To ensure that children could see the researcher and the study images, participants first completed a screen check. For children younger than 7 years of age, after sharing their screen, the

researcher made a thumbs-up sign and asked children if they could "do what I'm doing with my hand." Then, children saw a picture of a blue star and a red circle (see **Figure 1**, top) and were asked to name the color of each shape. Children were asked which shape was bigger if they could not name the colors (e.g., one parent reported that their child was colorblind). For studies of children age 7 years and older, the researcher first held up three fingers and asked the child "how many fingers am I holding up?" Next, they asked the child to hold up two fingers. Finally, an image with five shapes was shown, and the child was asked the color of the rectangle and diamond. If they could not name the colors, the child was asked how many shapes they saw (see **Figure 1**, bottom).

General Study Procedure

After the screen-sharing check, across research questions our studies involved showing children pictures of people and/or foods. Some studies included short stories about characters featured in the studies. Then, we asked children questions about these pictures or stories. For example, in one study, we showed children pictures about foods from different cultures, asked them their opinions of each of those foods, and who they think would be more likely to bring that food to school from an array of faces (Venkatesh and DeJesus, 2021). In another study, we showed children stories about characters who were sick and asked them to make predictions about disease transmission (DeJesus et al., in press).

Results

Among younger children, all children passed the thumbs-up check and all passed a version of the shape check (92 passed the color check and eight passed the size comparison check). Among older children, all children passed the holding-up fingers check, and all passed the shape check (one child only answered the color of the diamond).

Across our studies, our *a priori* exclusion criteria were as follows:

- (1) The child cannot see the researcher's screen or experienced Internet connectivity issues ($n = 1$).
- (2) The child asks to stop the study or walks away from the screen without intention of returning to the study ($n = 4$).
- (3) The child observes their sibling participate before them or their sibling interferes with the study ($n = 1$).
- (4) We do not receive the parent online consent form ($n = 1$).
- (5) The parent interferes with the study ($n = 3$), and
- (6) The parent signed up, but child was not of the correct age for the study ($n = 1$).

Note that parent interference was defined by a parent suggesting an answer or commenting on the child's answer (such as "you like taking sandwiches to school!"). Responses were not excluded if the parent reminded the child to answer the researcher's questions but without suggesting what the answer should be (such as "look, she is asking you a question!"); directing the child's attention back to the researcher was

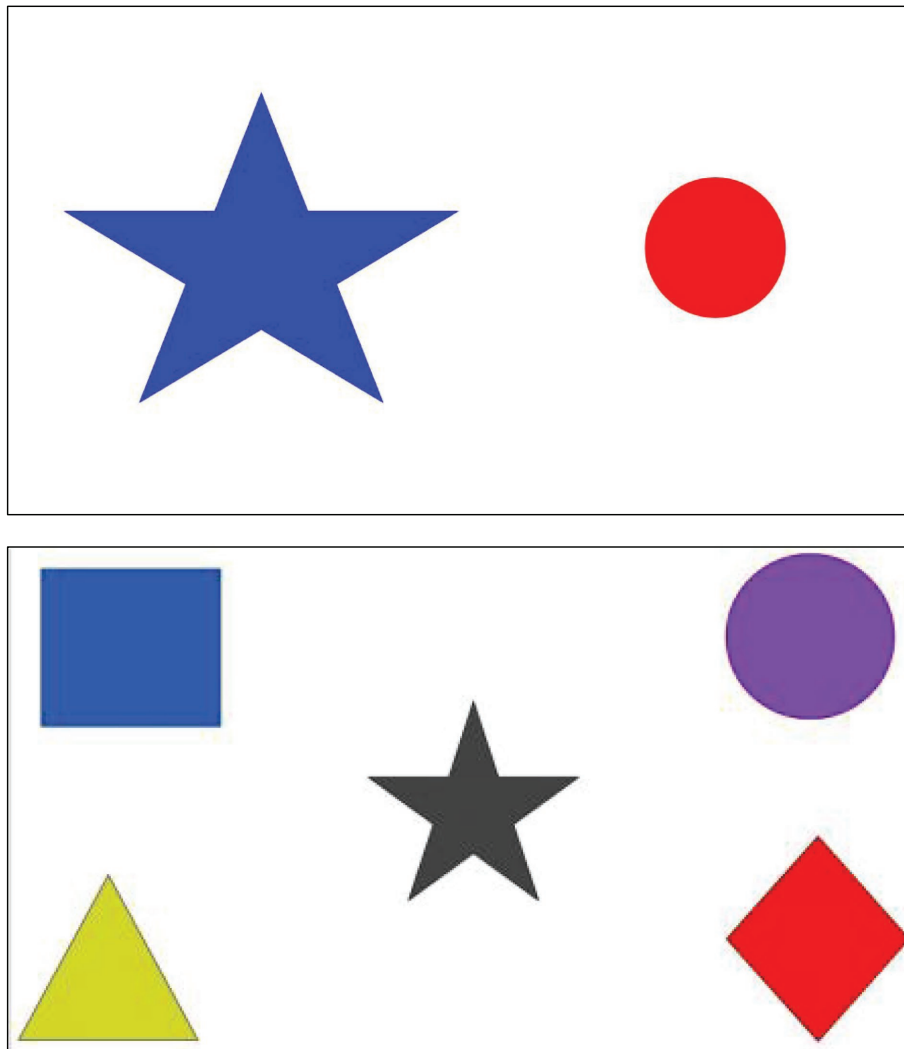


FIGURE 1 | Images shown to children (top: under 7 years; bottom: 7 years and older).

especially helpful for studies with younger children (3- and 4-year-olds).

Discussion

We had low rates of exclusion in studies of children's preferences and predictions *via* synchronous videoconference sessions. We excluded 11 children and retained data for 181 participants (94%). From our experience as researchers, children's ability to complete the session and share their opinions and preferences seemed comparable to in-person studies that are similar in format to the method described here. In line with our subjective experience, in a study that compared children's thinking about disease transmission in person before the pandemic and on Zoom during the pandemic, we found little difference in children's responses across time and platform (DeJesus et al., *in press*). Although we had few exclusions in these studies, anecdotally, studies that involved telling stories to children and asking them follow-up questions were especially challenging

for children younger than age 4. We return to this issue in the General Discussion.

GENERAL DISCUSSION

In this paper, we have illustrated two ways to study children's eating behavior at home using synchronous videoconference sessions. In the first method, we highlight a feasibility study in which we remotely observed meals and snacks at home with children under 3 years of age. Our analyses reveal that such designs yield video data that can be used for behavioral coding projects, based on the clear view of the feeding setup, child's face, and parent-child engagement in most videos. The main benefit of this paradigm is its ecological validity. Studies of eating behavior that are primarily conducted in settings outside the child's home, such as in the laboratory or in structured observations at schools or community centers that resemble in-lab studies (Fernandez et al., 2018; DeJesus et al.,

2018b), are valuable but may not be representative of the child's typical food environment. Our observational study which tested children at home provides a method to study children's eating behavior in a familiar environment. Children ate at their typical seats using cutlery and utensils they were familiar with, which may be especially useful to study children's reactions to familiar vs. unfamiliar foods (Moding et al., 2014; DeCosta et al., 2017). Testing children at home removes the additional variable of the unfamiliarity of an in-lab setting.

From a logistical perspective, studying children's eating behavior at home reduces the personnel and setup required for in-person lab feeding studies. First, in-person lab studies require a laboratory space, ideally with parking and access to public transportation, which researchers may not have available to them. Then to offer foods in an in-person lab study, researchers face additional challenges, including acquiring foods (especially for researchers interested in studying children's willingness to eat vegetables and other perishable foods) and avoiding common allergens. Moreover, laboratory studies typically standardize foods across participants, yet a food that is unfamiliar to one child might be familiar to another. Thus, while researchers might lose control over the standardization of the foods and environments that are possible for in-lab studies, at-home observational studies give parents the option of choosing foods that are familiar or unfamiliar to their child. This approach also gives parents the flexibility to schedule the testing session according to the child's current meal schedule (especially for infants when their mealtimes are more variable) without having to travel to another location. Even for observational studies of children's eating behavior at home described previously, researchers face logistical hurdles in terms of making trips to families' homes. This may require researchers to have access to transportation (e.g., to directly observe families or pick up and drop off recording equipment) and requires parents to be comfortable inviting researchers into their homes.

Another advantage of synchronous videoconference sessions is the option of giving live feedback to the parents. This feedback can serve multiple purposes. First, researchers can provide instructions to improve data quality. Synchronous videoconference sessions allow researchers to guide parents to ensure the camera is positioned accurately (compared to distributing video cameras for parents to use at home). Second, researchers can use this feedback to give parents specific instructions for an experimental manipulation. Although we chose not to give parents any specific instructions to make the session as easy as possible for parents and assess whether videoconference would be a suitable platform for research measuring children's eating behavior, many types of specific instructions could be given to bring in some of the control of laboratory studies. For instance, researchers can tell parents what type of foods to feed their child, instruct parents with specific prompts (such as feed your child an unfamiliar food for 5 min), or provide standardized types and amounts of foods (e.g., through delivering foods directly to parents) depending on the research question at hand.

In Method 2, we were able to collect behavioral data from 3- to 12-year-old children on their ratings and predictions.

We had low rates of exclusion across studies (we were able to retain 94% of participants), and children were able to see our pictures and hear us accurately, as indicated by the screen-share check questions. Such methods closely resemble food preference and rating studies conducted in the laboratory or other community settings (Rioux et al., 2016; DeJesus et al., 2019a; Echelbarger et al., 2020). Studies were run directly from Qualtrics, which reduced the extra step of running the study on another platform (such as Microsoft PowerPoint) and entering the data separately. Qualtrics is limited in its video storage capacity, so studies that include showing videos to participants require alternative presentation methods (e.g., embedding YouTube videos in Qualtrics or showing the video from another platform). None of the studies described here include videos, so we do not have data on potential exclusions due to insufficient connectivity to play videos (either from the researcher's side or the participant's side), which would be more prone to disruption. However, Method 2 appeared to be especially difficult for children younger than 4 years of age, especially without videos or detailed animations. Although we did not collect systematic data on this experience, anecdotally, it was much more difficult to complete synchronous videoconference sessions with children younger than age 4 (and even for some 4-year-old children) in terms of their understanding of their interaction with the researcher. For instance, some parents reported that their child might not fully understand that they were interacting with a real person.

Limitations and Challenges

While the present research highlights the potential to use synchronous videoconference sessions to conduct research on children's eating behavior, we interpret our claims with caution. This method limits the types of measurements that researchers can include in their data to those that can be seen or heard. Many studies that measure children's eating behavior includes child body mass index (BMI) or infant weight-for-length z-scores as predictors or outcomes in their analyses (e.g., Bergmeier et al., 2015a; Lumeng et al., 2020; Ventura and Hupp, 2020), which cannot be measured directly in a videoconference session. One approach to estimating this data could be to use coding tools that just require still images from the videoconference sessions. For example, the Shapecoder tool was designed to provide a coding system for child BMI and has both high inter-rater reliability and is correlated with child BMI measurements (Park et al., 2018). Researchers interested in using this tool may need multiple unobscured angles of the child (i.e., not blocked by a table). A similar tool is not currently available for infants, but researchers could consider asking parents for their child's measurements at their last pediatrician visit. Although parents tend to underestimate their child's weight (Eckstein et al., 2006; Lundahl et al., 2014), parents may have access to this data electronically through their healthcare provider, or parents of infants could have better recollection for their infants' measurements due to more frequent pediatrician visits. We did not attempt to study the feasibility of collecting height and weight measurements in these studies, but it is possible that some estimate could be attainable.

Importantly, our participant demographics represent homogenous families who were majority White, highly educated, and from higher income brackets. We relied on the platform Children Helping Science for recruitment, which is frequented by parents who are researchers/faculty themselves and may be familiar with online research and the challenges of continuing research programs during the pandemic. The vast majority (85%) of our meal recordings did not have substantial Internet connectivity issues, and we excluded only one of our verbal preference studies for network connectivity disruptions. Our sample's higher socioeconomic status is suggestive of their access to stable Internet connections and technology (e.g., updated and reliable smartphones, tablets, or computers) which enabled them to participate in such studies. Although some note remote online testing as an opportunity to include families from diverse backgrounds in child development research (e.g., Rhodes et al., 2020), the digital divide may further exclude participants from minority and lower socioeconomic backgrounds who not only have limited access to the Internet connectivity required for online data collection, but who are also faced with economic and childcare inequalities and have been most impacted by the COVID-19 pandemic (Lourenco and Tasimi, 2020). Especially pertinent to food-related research, such populations are also more likely to encounter food insecurity and rely on food assistance programs during the pandemic (Gassman-Pines and Gennetian, 2020). Thus, while synchronous videoconference sessions allowed us to interact with families who were diverse geographically (rather than being limited to our local area), our sample is restricted in its racial/ethnic and socioeconomic diversity. Our feasibility findings can only be generalized to families who are from the similar social and economic backgrounds as in our sample. Similar concerns surrounding access also apply to our research team – our research assistants who previously assisted with in-person lab studies also needed sufficient technology and private spaces to assist with research studies by videoconference, potentially leading to inequities in access to high impact teaching practices, such as participating in hands-on research activities (e.g., Kuh, 2008).

This videoconference method required basic parental literacy of video-chat applications (i.e., being able to be seen and heard on video) that we also shared *via* a guide with them. We did not experience issues with the setup of the call in any of our sessions. While parents might be more familiar with certain video chat applications (such as FaceTime), Zoom, and WebEx provide the option to record to the device (and not the cloud) which enhances the safety of the recordings and provides a standard option across families (e.g., families that did not have Apple devices did not have access to FaceTime when we began the study). Anecdotally, with the ubiquitous use of Zoom during the year of the pandemic, parents and children were more familiar and comfortable with the application compared to when we initially used WebEx for data collection. Nonetheless, more research is needed to better describe children's understanding of interactions by video and their views on being videotaped, which may vary across children. Outside of our specific research questions, even young children are able to have positive interactions that build relational connections

on video (McClure and Barr, 2017; McClure et al., 2021), though this may not fully generalize to conversations with unfamiliar researchers they are meeting for one session. At the same time, while children's understanding of some aspects of digital privacy is developing (Gelman et al., 2018; Sun et al., 2021), more research is needed to better understand children's beliefs, knowledge, and preferences in this area.

Recommendations and Future Directions

Based on our experiences of conducting the present research, we have the following recommendations to researchers who seek to use synchronous videoconference sessions to study children's eating behavior:

- (1) Closely consider what data you hope to attain and develop instructions to ensure that behavior is visible on the video.
- (2) Plan on changing the requirements of those instructions based on the device the parent logs-in from. Different devices (smartphones, tablets, and laptops) contain varying ranges of view for a video frame, so consider asking the parent what device they are using and share instructions for positioning/lighting accordingly.
- (3) Studies that ask children to follow stories may not be accessible to children under the age of 4 or 5. There are many potentially interesting questions to ask with 2- to 4-year-olds that primarily observe children's behavior or enlist parents as the experimenter (rather than relying on their ability to interact on videoconference with an unfamiliar person).
- (4) Consider creating a demonstration version of your study in case of serious Internet connectivity issues. For instance, if families do not have sufficient Internet connectivity to pass the screen-sharing check or turn on their video, it will be helpful to have some open-ended questions for the child or parent to answer. Such demonstrations may be familiar to researchers who work in museums or other community settings, where it is often useful to have a related demonstration activity for children whose parent/guardian is not present or would prefer not to sign consent documents. This demonstration would still give families the opportunity to engage with the research process and discuss their experiences with the researcher.
- (5) Make use of the live session to ensure parents fill out the online consent form (if they have not already) before you start the session with the child and to clarify data entered in the consent form that may contain typos (for example, birthdates).
- (6) If possible, have Internet hot spots and additional technology available for members of the research team to check out. Note that hot spots may not improve Internet access in low coverage areas.
- (7) Target multiple social media and online platforms for recruitment. In addition, consider physical advertisements in your community. This may raise the profile of your research to families who may not be as reachable using social media.

In addition to these recommendations, there are several topics that we view as possible to study using remote methods but have not yet pursued. We review two here in more detail. First, before infants begin eating solid foods, food intake is often measured using sucking behavior. Although sucking behavior can be coded from video recordings (Lumeng et al., 2007), this may be a difficult task to complete over videoconference. Based on our small number of bottle feedings, detailed instructions for parents on camera placement would be needed to achieve the close and unobstructed view of the infant's face that is needed for video coding. Alternatively, devices, such as the Neonur, can record infants' continuous negative sucking pressure and sucking bursts, or clusters of sucks that occur within less than 2 s between each suck (Lumeng et al., 2020); however, such devices would need to be exchanged with parents (which may be challenging with limited interaction and available team members). Second, digital imaging can be used to identify foods on a plate and measure food intake. In an intervention that explored whether involving children in making foods would increase their willingness to try new foods, researchers assessed children's snack choices after the intervention by comparing pictures of their plates before and after intake (Allirot et al., 2016). Similarly, the contents and nutritional quality of children's packed lunches were coded from photographs of the participating children's lunchbox contents before children ate lunch (Sutter et al., 2019). Researchers can also measure the healthfulness of meals consumed through "plate analysis" or examining what types of foods are on children's plates, for instance using the Healthy Meal Index (Kasper et al., 2016). Parents could share pictures of children's plates/meals for analysis by researchers, an even smaller commitment of time and technology compared to a videoconference study.

CONCLUSION

This paper has illustrated how synchronous videoconference sessions can be used to study children's feeding behaviors, adding to existing work that uses these designs to examine children's cognition, emotion, language, and social development. Using these sessions to observe meals provides ecological validity for children's eating behaviors and allows for live researcher feedback. Various measures can be collected through these methods, such as bites or pieces eaten, meal characteristics (such as the feeding setup or whether it is a family meal), and parent-child talk during meals. While researchers may have to compromise the standardization of foods and environment that laboratory settings offer, we gain the generalizability of findings and increased

participant scheduling flexibility. Moreover, researchers can use videoconference sessions to verbally assess children's beliefs and preferences of different foods. While we are grateful for platforms, such as Children Helping Science, that have significantly enabled our laboratory's continued data collection during the pandemic, we are also mindful of the representation in our sample. Ultimately, there is much to be gleaned about children's eating behaviors and synchronous videoconference sessions can be a useful tool for researchers interested in connecting with families at home.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: Open Science Framework, <https://osf.io/rhmuq/>

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by UNC Greensboro Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SV and JD conceptualized the studies, collected the data for Methods 1 and 2, and contributed to writing the manuscript. SV directed the video issues coding and analyzed the data in Method 1. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We thank Todd Ross, Evelyn Marin-Sanchez, and Jennifer McFarland for coding and the Parent Researcher Collaborative (including Elizabeth Bonawitz, Hyowon Gweon, Julian Jara Ettinger, Candice Mills, Laura Schulz, and Mark Sheskin) for recruitment assistance through the Children Helping Science platform. We especially thank the families who participated in this study during the COVID-19 pandemic. We thank Shaylene Nancekivell for helpful conversation about this project and online data collection in general.

REFERENCES

- Allirot, X., da Quinta, N., Chokupermal, K., and Urdaneta, E. (2016). Involving children in cooking activities: a potential strategy for directing food choices toward novel foods containing vegetables. *Appetite* 103, 275–285. doi: 10.1016/j.appet.2016.04.031
- Bergmeier, H. J., Skouteris, H., Haycraft, E., Haines, J., and Hooley, M. (2015a). Reported and observed controlling feeding practices predict child eating behavior after 12 months. *J. Nutr.* 145, 1311–1316. doi: 10.3945/jn.114.206268
- Bergmeier, H., Skouteris, H., and Hetherington, M. (2015b). Systematic research review of observational approaches used to evaluate mother-child mealtime interactions during preschool years. *Am. J. Clin. Nutr.* 101, 7–15. doi: 10.3945/ajcn.114.092114
- Carnell, S., and Wardle, J. (2007). Measuring behavioural susceptibility to obesity: validation of the child eating behaviour questionnaire. *Appetite* 48, 104–113. doi: 10.1016/j.appet.2006.07.075
- Cruwys, T., Bevelander, K. E., and Hermans, R. C. (2015). Social modeling of eating: A review of when and why social influence

- affects food intake and choice. *Appetite* 86, 3–18. doi: 10.1016/j.appet.2014.08.035
- De Castro, J. M. (1994). Family and friends produce greater social facilitation of food intake than other companions. *Physiol. Behav.* 56, 445–455. doi: 10.1016/0031-9384(94)90286-0
- DeCosta, P., Møller, P., Frøst, M. B., and Olsen, A. (2017). Changing children's eating behaviour—A review of experimental research. *Appetite* 113, 327–357. doi: 10.1016/j.appet.2017.03.004
- DeJesus, J. M., Gelman, S. A., Herold, I., and Lumeng, J. C. (2019a). Children eat more food when they prepare it themselves. *Appetite* 133, 305–312. doi: 10.1016/j.appet.2018.11.006
- DeJesus, J. M., Gelman, S. A., Viechnicki, G. B., Appugliese, D. P., Miller, A. L., Rosenblum, K. L., et al. (2018a). An investigation of maternal food intake and maternal food talk as predictors of child food intake. *Appetite* 127, 356–363. doi: 10.1016/j.appet.2018.04.018
- DeJesus, J. M., Gerdin, E., Sullivan, K. R., and Kinzler, K. D. (2019b). Children judge others based on their food choices. *J. Exp. Child Psychol.* 179, 143–161. doi: 10.1016/j.jecp.2018.10.009
- DeJesus, J. M., Shutts, K., and Kinzler, K. D. (2018b). Mere social knowledge impacts children's consumption and categorization of foods. *Dev. Sci.* 21:e12627. doi: 10.1111/desc.12627
- DeJesus, J. M., and Venkatesh, S. (2020). Show or tell: Children's learning about food from action vs verbal testimony. *Pediatr. Obes.* 15:e12719. doi: 10.1111/ijpo.12719
- DeJesus, J. M., Venkatesh, S., and Kinzler, K. D. (in press). Young children's ability to make predictions about novel illnesses. *Child Dev.*
- Echelbarger, M., Maimaran, M., and Gelman, S. A. (2020). Children's variety seeking in food choices. *J. Assoc. Consum. Res.* 5, 322–328. doi: 10.1086/709172
- Eckstein, K. C., Mikhail, L. M., Ariza, A. J., Thomson, J. S., Millard, S. C., and Binns, H. J. (2006). Parents' perceptions of their child's weight and health. *Pediatrics* 117, 681–690. doi: 10.1542/peds.2005-0910
- Edelson, L. R., Mokdad, C., and Martin, N. (2016). Prompts to eat novel and familiar fruits and vegetables in families with 1–3 year-old children: relationships with food acceptance and intake. *Appetite* 99, 138–148. doi: 10.1016/j.appet.2016.01.015
- Fernandez, C., DeJesus, J. M., Miller, A. L., Appugliese, D. P., Rosenblum, K. L., Lumeng, J. C., et al. (2018). Selective eating behaviors in children: An observational validation of parental report measures. *Appetite* 127, 163–170. doi: 10.1016/j.appet.2018.04.028
- Gassman-Pines, A., and Gennetian, L. A. (2020). COVID-19 job and income loss jeopardize child well-being: income support policies can help. *Soc. Res. Child Dev.* 9, 1–2.
- Gelman, S. A., Martinez, M., Davidson, N. S., and Noles, N. S. (2018). Developing digital privacy: Children's moral judgments concerning mobile GPS devices. *Child Dev.* 89, 17–26. doi: 10.1111/cdev.12826
- Gripshover, S. J., and Markman, E. M. (2013). Teaching young children a theory of nutrition: conceptual change and the potential for increased vegetable consumption. *Psychol. Sci.* 24, 1541–1553. doi: 10.1177/0956797612474827
- Kasper, N., Mandell, C., Ball, S., Miller, A. L., Lumeng, J., and Peterson, K. E. (2016). The healthy meal index: a tool for measuring the healthfulness of meals served to children. *Appetite* 103, 54–63. doi: 10.1016/j.appet.2016.02.160
- Kuh, G. D. (2008). Excerpt from high-impact educational practices: what they are, who has access to them, and why they matter. *Assoc. Am. Coll. Univ.* 14, 28–29.
- Kuo, P., Sayfer, P., Warmuth, K. A., Laifer, L., Parent, J., and McDaniel, B. T. (2021). “Practices for obtaining high-quality data from families online [Professional Development Session],” in *Biennial Meeting of the Society for Research in Child Development*; April 8, 2021.
- Lieberman, Z., Woodward, A. L., Sullivan, K. R., and Kinzler, K. D. (2016). Early emerging system for reasoning about the social nature of food. *Proc. Natl. Acad. Sci.* 113, 9480–9485. doi: 10.1073/pnas.1605456113
- Lourenco, S. F., and Tasimi, a. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Lumeng, J. C., Patil, N., and Blass, E. M. (2007). Social influences on formula intake via suckling in 7 to 14-week-old-infants. *Dev. Psychobiol.* 49, 351–361. doi: 10.1002/dev.20221
- Lumeng, J. C., Weeks, H. M., Asta, K., Sturza, J., Kaciroti, N. A., Miller, A. L., et al. (2020). Sucking behavior in typical and challenging feedings in association with weight gain from birth to 4 months in full-term infants. *Appetite* 153:104745. doi: 10.1016/j.appet.2020.104745
- Lundahl, A., Kidwell, K. M., and Nelson, T. D. (2014). Parental underestimates of child weight: a meta-analysis. *Pediatrics* 133, e689–e703. doi: 10.1542/peds.2013-2690
- Maimaran, M., and Fishbach, A. (2014). If it's useful and you know it, do you eat? Preschoolers refrain from instrumental food. *J. Consum. Res.* 41, 642–655. doi: 10.1086/677224
- McClure, E., and Barr, R. (2017). “Building family relationships from a distance: supporting connections with babies and toddlers using video and video chat,” in *Media Exposure During Infancy and Early Childhood*. eds. R. Barr and D. N. Linebarger (Cham: Springer), 227–248.
- McClure, E., Blanchfield, O., Myers, L. J., Roche, E., Strouse, G. A., Stuckelman, Z., et al. (2021). “Zoom-ing through development: Using video chat to support family connections during COVID-19,” in *Is Screen Time Family Time? Media in the Family System During COVID-19: Biennial Meeting of the Society for Research in Child Development* (Symposium presentation). ed. L. J. Myers; April 9, 2021.
- Mennella, J. A., Jagnow, C. P., and Beauchamp, G. K. (2001). Prenatal and postnatal flavor learning by human infants. *Pediatrics* 107:E88. doi: 10.1542/peds.107.6.e88
- Moding, K. J., Birch, L. L., and Stifter, C. A. (2014). Infant temperament and feeding history predict infants' responses to novel foods. *Appetite* 83, 218–225. doi: 10.1016/j.appet.2014.08.030
- Musher-Eizenman, D., and Holub, S. (2007). Comprehensive feeding practices questionnaire: validation of a new measure of parental feeding practices. *J. Pediatr. Psychol.* 32, 960–972. doi: 10.1093/jpepsy/jsm037
- Norton, M. I., Mochon, D., and Ariely, D. (2012). The IKEA effect: when labor leads to love. *J. Consum. Psychol.* 22, 453–460. doi: 10.1016/j.jcps.2011.08.002
- Park, B. K. D., Reed, M. P., Kaciroti, N., Love, M., Miller, A. L., Appugliese, D. P., et al. (2018). SHAPECODER: a new method for visual quantification of body mass index in young children. *Pediatr. Obes.* 13, 88–93. doi: 10.1111/ijpo.12202
- Raghoobar, S., van Kleef, E., and de Vet, E. (2017). Self-crafting vegetable snacks: testing the IKEA-effect in children. *Br. Food J.* 119, 1301–1312. doi: 10.1108/BFJ-09-2016-0443
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Rioux, C., Picard, D., and Lafraire, J. (2016). Food rejection and the development of food categorization in young children. *Cogn. Dev.* 40, 163–177. doi: 10.1016/j.cogdev.2016.09.003
- Scott, K., and Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Semmelmann, K., Hönekopp, A., and Weigelt, S. (2017). Looking tasks online: utilizing webcams to collect video data from home. *Front. Psychol.* 8. doi: 10.3389/fpsyg.2017.01582
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Shutts, K., Kinzler, K. D., McKee, C. B., and Spelke, E. S. (2009). Social information guides infants' selection of foods. *J. Cognit. Dev.* 10, 1–17. doi: 10.1080/15248370902966636
- Sun, K., Sugatan, C., Afnan, T., Simon, H., Gelman, S. A., Radesky, J., et al. (2021). “They see you're a girl if you pick a pink robot with a skirt: a qualitative study of how children conceptualize data processing and digital privacy risks,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*; May 8–13, 2021; 1–34.
- Sutter, C., Taylor, J. C., Nishina, A., and Ontai, L. L. (2019). Parental and family predictors of fruits and vegetables in elementary school children's home-packed lunches across a school week. *Appetite* 133, 423–432. doi: 10.1016/j.appet.2018.12.003
- Tran, M., Cabral, L., Patel, R., and Cusack, R. (2017). Online recruitment and testing of infants with mechanical Turk. *J. Exp. Child Psychol.* 156, 168–178. doi: 10.1016/j.jecp.2016.12.003

- Venkatesh, S., and DeJesus, J. M. (2021). "What's in your Dabba? Children's evaluations of ethnic lunchbox foods," in *What's in a Group? New Directions for Children's Essentialist Perceptions: Biennial Meeting of the Society for Research in Child Development (Symposium presentation)*; ed. B. Straka; April 8, 2021.
- Ventura, A., and Hupp, M. (2020). A within-subject comparison of maternal sensitivity to infant cues and infant intake during breastfeeding versus bottle-feeding interactions. *Curr. Dev. Nutr.* 4(Suppl. 2):1094. doi: 10.1093/cdn/nzaa054_166
- Ventura, A. K., Li, R., and Xu, X. (2020). Associations between bottle-feeding during infancy and obesity at age 6 years are mediated by greater infancy weight gain. *Child. Obes.* 16, 316–326. doi: 10.1089/chi.2019.0299
- Wardle, J., Guthrie, C. A., Sanderson, S., and Rapoport, L. (2001). Development of the Children's eating behaviour questionnaire. *J. Child Psychol. Psychiatry* 42, 963–970. doi: 10.1111/1469-7610.00792
- Zeinstra, G. G., Renes, R. J., Koelen, M. A., Kok, F. J., and de Graaf, C. (2010). Offering choice and its effect on Dutch children's liking and consumption of vegetables: a randomized controlled trial. *Am. J. Clin. Nutr.* 91, 349–356. doi: 10.3945/ajcn.2009.28529

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Venkatesh and DeJesus. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Research at a Distance: Replicating Semantic Differentiation Effects Using Remote Data Collection With Children Participants

Catarina Vales^{1*}, Christine Wu¹, Jennifer Torrance², Heather Shannon², Sarah L. States² and Anna V. Fisher¹

¹ Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, United States, ² Phipps Conservatory and Botanical Gardens, Pittsburgh, PA, United States

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Yang Wu,
Stanford University, United States
Kevin Darby,
University of Virginia, United States

*Correspondence:

Catarina Vales
cvales@andrew.cmu.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 19 April 2021

Accepted: 30 June 2021

Published: 06 August 2021

Citation:

Vales C, Wu C, Torrance J,
Shannon H, States SL and Fisher AV
(2021) Research at a Distance:
Replicating Semantic Differentiation
Effects Using Remote Data Collection
With Children Participants.
Front. Psychol. 12:697550.
doi: 10.3389/fpsyg.2021.697550

Remote data collection procedures can strengthen developmental science by addressing current limitations to in-person data collection and helping recruit more diverse and larger samples of participants. Thus, remote data collection opens an opportunity for more equitable and more replicable developmental science. However, it remains an open question whether remote data collection procedures with children participants produce results comparable to those obtained using in-person data collection. This knowledge is critical to integrate results across studies using different data collection procedures. We developed novel web-based versions of two tasks that have been used in prior work with 4-6-year-old children and recruited children who were participating in a virtual enrichment program. We report the first successful remote replication of two key experimental effects that speak to the emergence of structured semantic representations ($N = 52$) and their role in inferential reasoning ($N = 40$). We discuss the implications of these findings for using remote data collection with children participants, for maintaining research collaborations with community settings, and for strengthening methodological practices in developmental science.

Keywords: semantic structure, semantic differentiation, semantic similarity, spatial arrangement method, semantic inference, remote data collection

INTRODUCTION

The field of developmental science is in urgent need of assessing remote data collection procedures. The majority of data collection in developmental science – whether observational or experimental – has traditionally relied on in-person data collection. However, there is a growing recognition that in-person data collection procedures place barriers to participation from underrepresented populations and make large samples difficult to attain. More recently, limitations to in-person data collection resulting from public health mitigation strategies due to the COVID-19 pandemic further highlighted the need for developing and evaluating remote data collection procedures. Here we replicate two semantic differentiation effects that were previously documented in 4-6-year-old children using in-person data collection and report the extent to which these effects are robust to variation in testing conditions that are typically well controlled during in-person data collection. We also describe an efficient recruitment strategy – enrolling children participating in

virtual enrichment programs – that can allow researchers to broaden community partnerships. These findings point to the feasibility of conducting rapid, robust, and replicable research with children using remote data collection procedures.

Increasing Need for Remote Data Collection With Children Participants

In the United States, developmental science has historically relied on in-person data collection procedures. At the beginning of the 20th century, a number of university-affiliated laboratories dedicated to documenting children's development began the practice of inviting children and their caregivers to research facilities on campus to observe and assess behaviors of interest (Gesell, 1932; Ossmer, 2020). This recruitment strategy led to a number of important discoveries in the field, and is still used by many research labs to this day. However, because this approach requires participants to travel to the laboratory, it often results in study samples that are not only small (because this method is time-consuming) but also highly homogenous (because caregivers who have time and resources to travel to university laboratories come largely from White and mid- to high socioeconomic status communities). Small and homogeneous samples limit the conclusions that can be drawn from developmental studies for two reasons. First, the use of small sample sizes decreases statistical power. Statistical power is not only critical to detect true effects, but – at first glance, counterintuitively – low statistical power can *decrease* the likelihood that *significant effects* are indeed *true effects* (Button et al., 2013). In other words, the use of small sample sizes can lead to an increase of false positives. Second, homogenous samples obscure the impact of a multitude of variables on research findings, thus impeding both theoretical and empirical progress (Fernald, 2010; Henrich et al., 2010; Varga, 2011; Sugden and Moulson, 2015; Nielsen et al., 2017).

To address these concerns, researchers have developed community-based recruitment strategies that can facilitate the recruitment of larger and more diverse samples. For example, researchers have recruited and collected data in children's museums, after-school programs, pediatricians' offices, and mobile laboratories (e.g., Alibali and Nathan, 2010; Callanan, 2012; Cates et al., 2018). These approaches have been successful at increasing the size and diversifying study samples and are important methodological advances in the field of developmental science. However, these approaches are still limited by the geographical location of the recruitment sites and the make-up of the population they serve. For example, while recruiting participants at a children's museum can lead to the recruitment of samples that are larger and racially more diverse, admission fees to the museum may still be a barrier to recruiting economically diverse samples.

Remote data collection procedures have the potential to help recruit larger and more representative samples of participants – in regards to race and ethnicity, income, and geographical location of the participants – into developmental studies [Scott and Schulz, 2017; Sheskin and Keil, 2018; Rhodes et al., 2020; but see Lourenco and Tasimi (2020) for how researchers should

consider possible inequalities in internet access when planning remote studies]. In the last year, there was also increased interest in conducting research remotely as mitigation strategies in response to the COVID-19 pandemic severely limited the ability to collect data in person. Even with the onset of mass vaccination plans and as social distancing protocols are gradually relaxed, in-person data collection will likely not immediately return to the rates observed prior to the pandemic – making remote data collection procedures increasingly common in the coming years.

Despite the potential advantages and increased need of remote data collection procedures, and despite a number of recent studies using remote data collection with children participants (e.g., Chuey et al., 2020; Leshin et al., 2021), there is currently a gap in the *evaluation* of remote data collecting procedures used with children. It is thus critical to evaluate whether remote data collection procedures can assess constructs of interest in ways that are comparable to in-person data collection. If so, then developmental scientists can confidently use remote data collection procedures to continue to accumulate knowledge and integrate findings from remote studies with work conducted in-person.

It may seem trivial that children would perform equivalently on cognitive tasks regardless of whether they are assessed in person or remotely. Children in the United States are likely familiar with technology (Rideout, 2017; Chen and Adler, 2019), and many existing research protocols for in-person data collection are already computerized (e.g., Friend and Keplinger, 2003; Gershon et al., 2010; Fisher et al., 2013). Similarly, children are possibly more comfortable and thus more likely to engage with a task in a known setting such as their home (see Klein and Durfee, 1979; Belsky, 1980; Perry et al., 2014; Santolin et al., 2021 for related arguments). In sum, there are reasons to be optimistic about remote data collection procedures with children participants.

However, remote data collection procedures likely introduce additional variability in the setting and measurement that could limit the feasibility of remote data collection, particularly with young children. For example, while computerized assessments collected in-person standardize features such as the size of the screen used to display the task or the distance at which children sit from the screen, these factors will vary considerably when participants complete tasks remotely using their own devices. Additionally, it is also possible that children encounter more distractions when at home, that the absence of an experimenter next to the child to explain, scaffold, and redirect the child to the task when necessary, and that possible influences from caregivers would make data collection considerably less successful. Thus, it is important to ensure that – despite these potential sources of variability – data collected remotely with young children participants is comparable to data obtained from in-person assessments. While recent work has shown that remote data collection procedures can replicate the effects of lab-based studies in older children and adolescents (Nussenbaum et al., 2020), it remains an open question whether data collected remotely with young children is comparable to data obtained in-person.

Here, we address this goal by aiming to replicate two semantic differentiation effects that were previously observed

in 4–6-year-old children using in-person data collection (Fisher et al., 2015; Vales et al., 2020a,b). Using remote data collection procedures, we asked whether we could conceptually replicate these effects. We did not aim to collect a representative sample or obtain a sample size larger than in prior studies (although we ultimately enrolled a larger number of participants than prior studies); rather, the main goal of this study was to provide a proof-of-concept that remote data collection procedures can measure constructs of interest in ways that are comparable to in-person data collection.

Prior Work on Semantic Differentiation in Children

Measuring Semantic Differentiation Using the Spatial Arrangement Task

Organized semantic representations, linking words and the concepts to which they refer by relevant within- and across-domain distinctions, are believed to be a critical aspect of human cognition (Clark, 1973; Bjorklund and Jacobs, 1985; Gobbo and Chi, 1986). As such, there is a large interest in understanding how semantic structure develops with experience and learning, and how organized semantic representations influence other cognitive processes. Prior work suggests that children acquire structured semantic representations by exploiting the similarity structure of the entities in the world as they gradually learn about their features (Rogers and McClelland, 2004; Kemp and Tenenbaum, 2008; Hills et al., 2009). One aspect of many common domains in the world (e.g., animals, plants, clothes, tools, etc.) is that across-domain distinctions (e.g., animals vs. plants) rely on mostly non-overlapping clusters of features (e.g., only animals have eyes and can move, and only plants have leaves and roots), while within-domain distinctions (e.g., birds vs. mammals) rely on partially overlapping clusters of features (e.g., beaks and feathers vs. fur and nursing their young all overlap with the presence of eyes and mobility). This structure should lead to across-domain distinctions being generally more strongly represented earlier in development relative to within-domain distinctions.

Two recent studies directly tested this prediction using a spatial arrangement task (Goldstone, 1994) in which children were asked to arrange items by placing related items close together; the physical distance between item pairs served as a proxy for semantic relatedness, with items judged as more similar placed closer together. These studies showed that younger children (4–6 years-old) strongly differentiated items belonging to different domains – placing pairs of items of the same domain closer together relative to pairs of items of different domains (Vales et al., 2020a,b). Reliable within-domain distinctions were only visible in older children or after extended experience with a domain (Vales et al., 2020a,b).

Although prior work with adult participants has used computerized versions of the spatial arrangement method (e.g., Goldstone, 1994; Koch et al., 2020), the existing studies with children participants using this task asked children to organize physical cards on a game board (e.g., Fisher et al., 2015; Jenkins et al., 2015; Vales et al., 2020a,b). Thus, it remains

an open question whether a computerized version of the spatial arrangement task would result in patterns of semantic differentiation similar to those observed in prior work. Here, we implemented and tested the first child-friendly computerized version of the spatial arrangement method.

Measuring Semantic Differentiation Using the Semantic Inference Task

Organized semantic representations critically support other cognitive processes, including the ability to make inductive inferences – such as assuming that members of the same within-domain group are likely to share features (e.g., Gelman and Markman, 1986; Gobbo and Chi, 1986; Coley, 2012; Fisher et al., 2015). Inductive inferences are often tested with a forced-choice semantic inference task in which children are asked to extend a property from a target item to one of a number of alternatives; for example, children might be asked whether a ‘sheep’ or a ‘cow’ shares a non-obvious feature with a ‘lamb.’ Consistent with the idea that children rely on organized semantic representations to make choices in this task and that close semantic representations compete for selection, the likelihood that children select the strongest-related item in this task is modulated not only by the similarity between the target and the match (i.e., lamb-sheep), but also by the similarity between the target and *the lure* – children are more likely to select ‘sheep’ as a match to ‘lamb’ in the presence of ‘clock’ (a lure belonging to a different domain) than in the presence of ‘cow’ (a lure belonging to the same domain) (Fisher et al., 2015).

Prior work with children using match-to-sample procedures like the one used in the semantic inference task has employed a range of number of trials (e.g., Tversky, 1985; Waxman and Namy, 1997; Fisher et al., 2015). Increasing the total number of trials completed by each participant is a crucial way to increase the precision – and thus, the power – of a task’s measurement (Forrester, 2015; DeBolt et al., 2020), but increasing the number of trials comes at the cost of possible attrition. Here, we implemented and tested a child-friendly adaptive procedure in which children could decide whether to continue or end the semantic inference task at the end of each block of trials.

The Present Study

Together, the findings described above speak to the mechanisms supporting the acquisition of structured semantic representations and how such semantic representations support inductive inferences. The goal of this study was to conduct a conceptual replication of (1) the differences in representational strength between across- and within-domain differentiation and (2) the lure distance effect in semantic inference in 4- to 6-year-old children. If semantic structure can be assessed remotely, then one should observe similar results with a remote sample – (1) weaker representation of within-domain distinctions relative to across-domain distinctions as measured by the spatial arrangement task, and (2) lower likelihood of selecting a match in the presence of a close versus distant lure in the semantic inference task. Thus, the present study aims to conceptually replicate these two effects with remote data collection procedures.

To do so, we recruited a sample of 4- to 6-year-old children as this is the age range in which both of these experimental effects have been observed in prior work. Children participants were enrolled in an out-of-school enrichment program – aiming to provide children with hands-on, educational activities – delivered remotely by a science center. As part of the program, children completed the task on their web browser while connected in a video call with a researcher; although data collection was not fully unmoderated (cf. Rhodes et al., 2020) as caregivers were not always available during the virtual program, the tasks were set up to require minimal interaction with the researcher – all the instructions and transitions between the protocol steps were interactively delivered in the browser.

The present study also aims to extend prior work examining the relation between semantic differentiation and inductive inferences. Consistent with the idea that children rely on organized semantic representations to make inductive inferences, the degree of a child's semantic differentiation appears to be related to their ability to make category-based inferences. Fisher et al. (2015) showed that a child's tendency to select a within-domain category match in the inductive inference task was positively associated with how strongly the child differentiated items within a domain. Children's within-domain semantic differentiation was assessed using the spatial arrangement method by comparing the distance at which category-matching (e.g., 'sheep') and habitat-matching (e.g., 'horse') items were placed from targets (e.g., 'lamb') – with larger distances indicating stronger differentiation. Children's inductive inferences were assessed using the semantic inference task by examining the likelihood of selecting a category-matching item (e.g., 'sheep') as having the same property as a target item (e.g., 'lamb'); importantly, as lure distance was not manipulated in this study, all lures in the inductive inference task were items that belonged to the same domain but not to the same category as the target (e.g., 'frog'). In the current study we will take advantage of collecting both semantic differentiation and inductive inference assessments to further examine this relation. Specifically, we will examine the relation between within-domain semantic differentiation and the likelihood of selecting a within-domain category match in the inference task. We note that there are a number of design differences between the current study and this prior work that may make the assessment of this association not trivial; we will return to this issue when discussing the findings of this analysis.

MATERIALS AND METHODS

Participants

We recruited a total of 58 children between 4 and 6 years of age who were enrolled in a week-long virtual enrichment program hosted by a botanical garden in Pittsburgh, PA, United States during the Summer of 2020; data were collected over three consecutive weeks, on a single day each week. To reduce economic barriers to participation, enrollment costs were partially waived. The caregiver-reported (provided to the botanical garden by 38 caregivers) gender and racial makeup of

the sample was 32% male, 63% female, and 5% not reported; 79% white, 5% Black/African American, 8% Asian/Indian American, and 8% multiracial. This sample was more racially diverse than Vales et al. (2020a), which recruited from the same botanical garden but during in-person enrichment programs (see **Supplementary Table S1**); we will return to this point in the "Discussion" section. The same caregivers also provided their zip code information; the majority of the participants ($N = 33$) lived in Pennsylvania, with 24 unique zip codes reported; the remaining participants lived in one of four states ($N = 4$) and in Canada ($N = 1$).

Data from six children were not recorded due to technical difficulties (unstable internet connection, $N = 4$; incompatible devices, $N = 2$) and were therefore not included in the analyses reported. Forty children completed both the spatial arrangement and the inference task, and 12 children completed the spatial arrangement task but not the inference task; thus, analyses of the spatial arrangement task include 52 participants and analyses of the inference task include 40 participants.

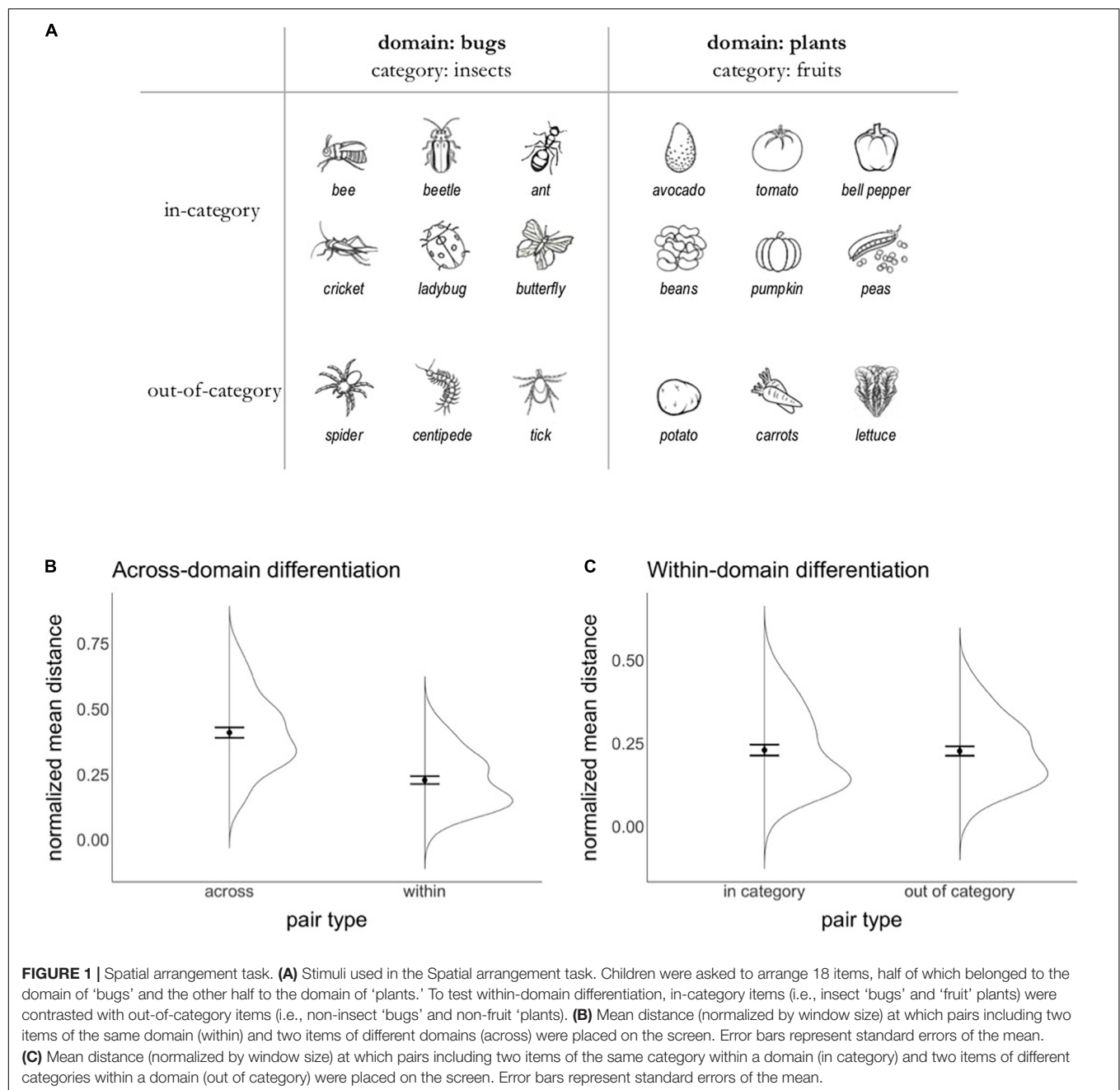
Children completed the tasks reported here before the start or during the first day of the enrichment program activities. Because these tasks were part of the enrichment program activities, in accordance with the IRB protocol approved by Carnegie Mellon University all children enrolled in the program were invited to complete the tasks. Caregivers were given the option to have their children's data excluded from analyses; no caregiver requested that their child's data be excluded.

Stimuli and Design

Spatial Arrangement Task

The stimuli used in the Spatial Arrangement task are shown in **Figure 1A** and were identical to the stimuli used in Vales et al. (2020a); a comparison between Vales et al. (2020a) and the current study's sample, task design, and results is available in **Supplementary Table S1**. To probe both within- and across-domain differentiation in a single trial, the stimulus set included two domains ('bugs' and 'plants') with a within-domain distinction ('bugs' that are insects vs. not; 'plants' that are fruits vs. not). Each pair of items was classified as either belonging to the same domain vs. not (i.e., whether it included any two bugs or two plants vs. one bug and one plant); this allowed us to probe across-domain differentiation. In addition, within-domain pairs were further classified as either belonging to the same within-domain group (e.g., *insect* bugs) or not (e.g., *non-insect* bugs); this allowed us to probe within-domain differentiation. Black and white line drawings representing each item were presented as individual cards with a white background against the screen's black background.

The task was hosted in the Qualtrics platform by adapting the procedure developed by Koch et al. (2020). The pixel width and height of the center of each item was recorded, as well as the pixel width and height available on each participant's screen; these coordinates were used to calculate the distance between all pairs of items on the screen and normalize them by each participant's maximum possible dissimilarity (i.e., the diagonal of the participant's screen).



Inference Task

Supplementary Table S2 shows all the linguistic stimuli used in the Inference task; a comparison between Fisher et al. (2015) and the current study’s sample, task design, and results is available in **Supplementary Table S1**. The stimulus set included six targets (all insect ‘bugs’), six matches (all insect ‘bugs’), six close lures (all non-insect ‘bugs’), six distant lures (all ‘plants’), and six novel biological properties (e.g., “vespanix cells”). To prevent children from responding based only on visually available features and to decrease overlap with the spatial arrangement task, the items in this task were not depicted and children were instead told that the items were hiding

behind trees, rocks, or grass (in blocks 1, 2, and 3, respectively) (see Fisher et al., 2015 for a similar approach).

To probe the effect of lure distance, in each block of trials each target (an insect ‘bug’) was paired with a match (another insect ‘bug’), a close lure (a non-insect ‘bug’) and a distant lure (a ‘plant’). There were a total of six targets per block, and thus a total of 12 trials per block.

Across blocks, each target was paired with a different match, lures, and property; each combination of target, match, and lures included a similar number of syllables and no overlapping word onsets. The location of the match was counterbalanced across the left and right side of the screen (with the additional constraint

that the match was not presented on the same side on more than three consecutive trials), so that at the end of each block of trials the match item was equally likely to occur on either side.

There were five additional trials designed to ensure that children understood and were engaged with the task. In these trials, the target and the match items were parent/offspring animal pairs and the distant lures were vehicles (e.g., target: 'kitty'; match: 'cat'; lure: 'bus'). Because the target and the match are strongly related to one another, and both are unrelated to the lure, if children understood and were engaged with the task they should reliably select the category match on these trials. Two of these trials were presented at the start of the task as familiarization trials; the other three trials were presented once in each block.

The task was hosted on the lab.js platform (Henninger et al., 2019) and embedded in Qualtrics so that the transition from the spatial arrangement task to the inference task was seamless. The participant's response on each trial (left vs. right selection) was recorded. The files used to run these tasks are openly available: <https://osf.io/67gtc/>.

Procedure

Children were individually tested by a trained experimenter in a breakout room in the Zoom communication platform (see **Figure 2A**). The experimenter started by establishing a rapport with the child; if a caregiver was present, the experimenter requested that they do not influence the child's responses. After this initial warm-up period, the experimenter helped the child share their screen so that the experimenter could see the child's screen and help with any experiment logistics throughout the session if needed (e.g., instructing a participant who seemed unsure how to continue); for the majority of participants no such help was needed after they started the tasks. Participants were then sent a link to the study through the Zoom messaging screen, which opened a web browser window where both tasks were completed. To ensure that the audio and video features of the browser were compatible with the study's platform, there was a brief video that participants were asked to play.

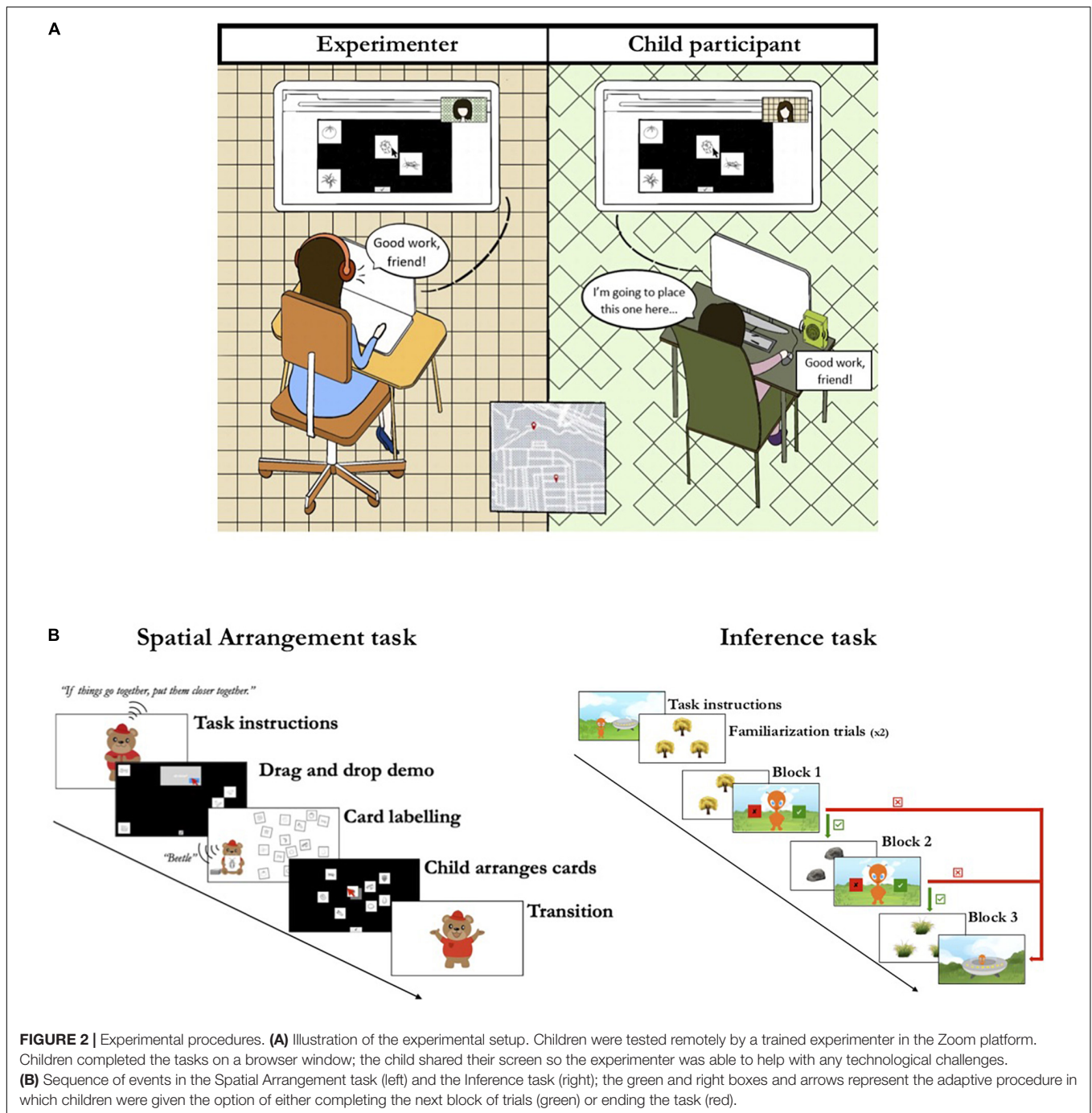
Once the audiovisual check was performed, children started the spatial arrangement task; **Figure 2B** shows the sequence of events in this task. An animated video narrated by a cartoon bear explained that the goal of the game was to organize cards on the screen by placing close together cards that go together, and place far apart cards that do not. Then the video transitioned to a tutorial of how to arrange the cards on a black screen by dragging and dropping them with the mouse; four cards displaying items unrelated to the study (a bus, a duck, a duckling, and a drum) were sorted by the bear. This part of the video displayed a computer screen with a visible mouse cursor and the bear's voice narrated while it walked through the task (e.g., "The bus does not go with the duck, so I will put them far apart"). The video ended with the bear character presenting and naming the cards that the child was asked to sort. The bear held one card at a time and labeled it (e.g., 'beetle'); after each card was labeled, it was added to the display of already-labeled cards floating on the screen beside the bear. The cards were

previewed in the same order by all children, and the labeled cards were not placed in a grid-like pattern so as to prevent biasing the child. After being shown all the cards to be sorted, children were instructed by the bear to press a button so they could start arranging their cards.

Once children advanced to the next screen, they were shown the screen where they would arrange the cards, a black background taking up the entirety of their browser window. Cards were presented one at a time in the center of the screen, in a random order for each participant. Children used their mouse, trackpad, or touchscreen to drag and drop each card anywhere on the screen. Once the first card was placed, a button at the bottom of the screen would become active and allow children to request the next card by clicking the button; this continued until all 18 cards were presented and arranged. After children arranged all cards, they were given a final opportunity to rearrange any cards before finishing the task. Upon completion, children were shown a transition video where they were thanked for their help and instructed to press a button when they were ready to start the second task.

Once children advanced to the second task, a video introducing the inference task started; **Figure 2B** shows the sequence of events in this task. Children were introduced to an alien and told that the goal of the game was to help the alien learn about animals and plants, which were hiding. On each trial, children were shown three identical objects (trees, rocks, or a patch of tall grass) arranged in an upright triangle pattern and were told the name of the organism hiding behind each object. For example, children heard something like: "There is a bee hiding behind this tree, a fruitfly hiding behind this tree, and a spider hiding behind this tree"; each object referred to was synchronously jittered to indicate the placement of each organism. The objects on the screen were always labeled and referred to in the same order: first the object on top, then the object on the bottom left side of the screen, followed by the object on the bottom right side of the screen. After being told which organism was hiding behind each object, children were then told that the target organism had a novel biological property (e.g., "The bee has drotium hairs") and asked to generalize this property to one of the two test organisms (e.g., "Which one also has drotium hairs?"); **Figure 3A** displays example trials. Children indicated their response by clicking on the item; only responses on the bottom left or right objects were accepted. Once children responded, the next trial started.

At the start of the task, after watching the introduction video, children were shown two familiarization trials that included a match and a distant lure from an unrelated domain (e.g., target: 'kitty,' match: 'cat,' lure: 'bus'); these trials were designed to present minimal competition between the match and the lure to make sure children understood the instructions. After these familiarization trials, children were shown three consecutive blocks of trials, each consisting of 12 test trials and 1 catch trial designed in a similar manner as the familiarization trials. After each block, children were given the option of continuing to the next block or ending the task. At the end of the task, a short video showed the alien thanking the child for their help and leaving Earth on a spaceship.



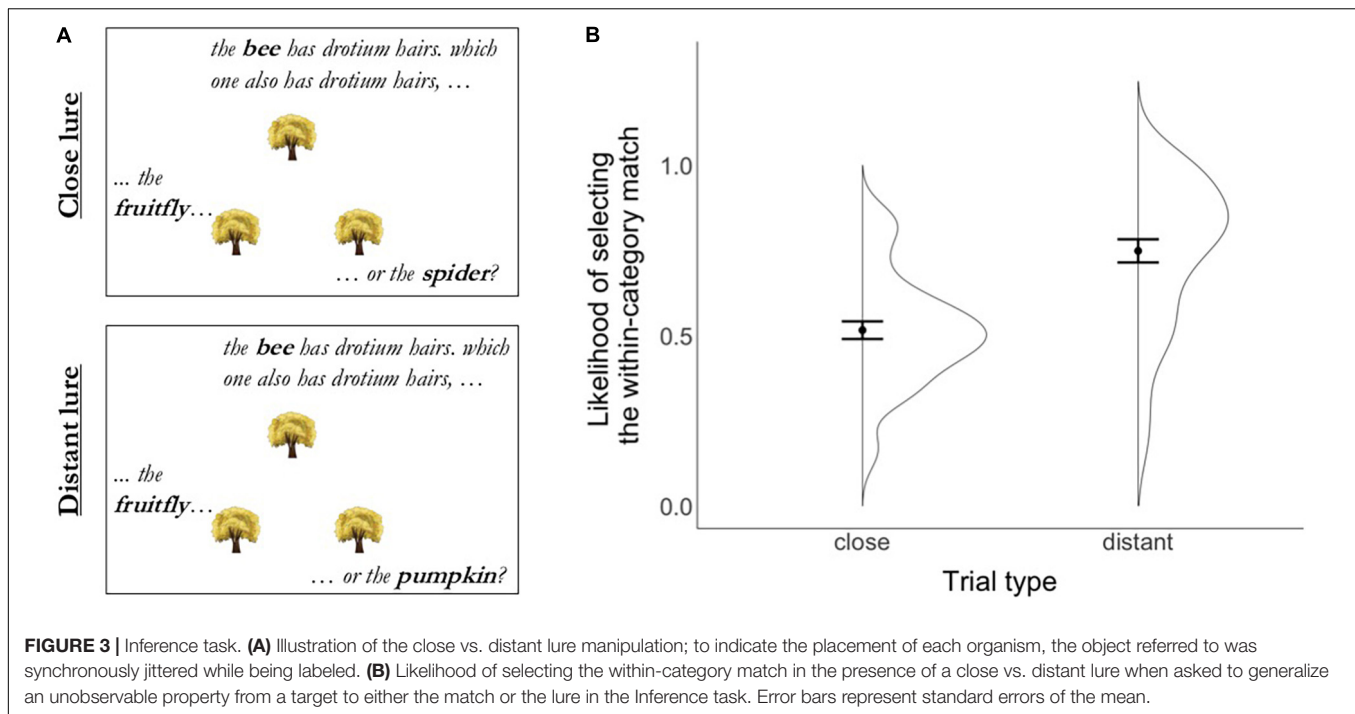
Once the child completed the second task, the experimenter thanked the child and any caregivers present and answered any questions they had. The child then rejoined the group activities taking place in the enrichment program.

RESULTS

We examined whether we could replicate previously reported differences in representational strength between across- and

within-domain differentiation (Vales et al., 2020a,b) and the lure distance effect in semantic inference (Fisher et al., 2015) using remote data collection procedures.

If an online version of the Spatial arrangement task, when delivered remotely, can provide estimates of semantic structure that are comparable to those obtained when children complete the task in person arranging physical cards on a game board, then we should see patterns of semantic differentiation similar to prior work (Vales et al., 2020a,b). Specifically, we would expect to see that children more strongly differentiate items belonging



to different domains of knowledge ('bugs' vs. 'plants') relative to items within a domain (i.e., insect vs. non-insect 'bugs'; fruit vs. non-fruit 'plants'). To examine this prediction, we compared the average distance at which children placed pairs including items of the same vs. different domains (to examine across-domain differentiation) and pairs including items of the same vs. different categories within a domain (to examine within-domain differentiation).

Similarly, if an online version of the Inference task, when delivered remotely, can provide estimates of inferential reasoning that are comparable to those obtained when children complete the task in person, then we should see a lure distance effect similar to prior work (Fisher et al., 2015). Specifically, we would expect to see a higher likelihood of extending a property from the target object to the match in the presence of a distant relative to a close lure. To examine this prediction, we compared the likelihood of selecting the match item in the presence of a close vs. distant lure.

To examine both of these predictions, we employed a linear mixed-effects approach to test the effect of the manipulation of interest on the outcome measure. Specifically, in the *Spatial arrangement task*, we tested the effect of pair type on the raw (i.e., non-averaged per participant) Euclidean distances between pairs of items. To account for differences in the space available to arrange the cards resulting from different sizes of browser windows, these pairwise distances were normalized (i.e., divided by the by the pixel length of the diagonal of the browser window; see Koch et al., 2020 for a similar approach). To examine whether using a larger browser window influenced children's likelihood of differentiating across or within domains, we included the size of the window in the models examining semantic differentiation. In the *Inference task* we tested the effect of lure type (close vs. distant) on the trial-by-trial likelihood of selecting the match

item. Because children were given the option to continue or end the task at the end of each block, we included the number of completed blocks in the model. For each of these predictions, we provide Cohen's *d* for the difference between the means of interest as a measure of effect size; as these predictions were tested with within-subjects manipulations, the correction suggested in Gibbons et al. (1993) was employed.

In addition to examining each task separately, we also examined the relation between the two tasks. Specifically, we examined whether the average degree of a child's within-domain differentiation (as measured by the Spatial Arrangement task) is predictive of a child's overall likelihood of selecting the match in the presence of the close lure in the Inference task.

Analyses were conducted in the R environment (R Core Team, 2014); except where noted we used the functions *lmer* and *glmer* from the 'lme4' package (Bates et al., 2015) to model continuous and binomial outcome variables, respectively. Variables were centered, with categorical variables coded using effects coding. Models were fit with the maximal random effects structure (Barr et al., 2013); we report model estimates for all models and *p*-values based on Wald tests of each model's fixed effects. The reported effect sizes were calculated with the function *cohen.d* from the 'effsize' package (Torchiano, 2020). Code and data are openly available: <https://osf.io/67gtc/>.

Spatial Arrangement Task

Figure 1B depicts the normalized average distance between pairs including two items from the same domain ('within') or from different domains ('across'), showing that children placed pairs of items belonging to the same domain closer together relative to pairs including items from different domains. A model testing the effect of pair type (within vs. across) and window

size confirmed that pair type was a significant predictor of the distance at which items were arranged on the screen [$b = -0.18$, $\chi^2(1) = 45$, $p < 0.0001$, Cohen's $d = 1.44$] but window size was not [$b = -0.000002$, $\chi^2(1) = 0.002$, $p = 0.97$]; the model included by-participant random intercepts and slopes for the effect of pair type. The effect size of the effect of pair type was of similar (albeit larger) magnitude relative to when data were collected in person (Vales et al., 2020a).

Figure 1C depicts the normalized average distance between pairs including two items from the same within-domain group ('in category') or from different groups ('out of category'), and shows that children placed the two types of pairs at similar distances. A model testing the effect of pair type (in vs. out of category) and window size showed that neither was a significant predictor of the distance at which items were arranged on the screen [pair type: $b = -0.003$, $\chi^2(1) = 0.36$, $p = 0.55$, Cohen's $d = -0.02$; window size: $b = 0.0007$, $\chi^2(1) = 1.11$, $p = 0.29$]; the model included by-participant random intercepts (the model including random slopes for the effect of pair type failed to converge). The effect size of the effect of pair type was of similar magnitude relative to when data were collected in person.

Together, these results provide a conceptual replication of prior work showing differences in representational strength between across- and within-domain differentiation (Vales et al., 2020a,b) using remote data collection procedures. The results also suggest that variation in the size of the web browser used to complete the spatial arrangement task is unlikely to contribute to children's degree of differentiation when completing the spatial arrangement task; in **Supplementary Material (Section C)** we present additional evidence that variation in the size of the browser window is not related to the degree of semantic differentiation (see **Supplementary Figure S1** in the **Supplementary Material**).

Inference Task

To ensure that children understood and were engaged with the Inference task, we started by examining their performance in the familiarization and catch trials. Children were highly accurate on both the familiarization trials at the beginning of the task ($M = 0.86$, $SD = 0.23$) and the catch trials interspersed among the test trials ($M = 0.85$, $SD = 0.23$), both significantly above chance (0.5) level [familiarization: $t(39) = 10.1$, $p < 0.0001$; catch: $t(39) = 6.96$, $p < 0.0001$]. Children were also likely to complete at least two blocks of test trials ($M = 2.25$, $SD = 0.86$), further suggesting that they were engaged with the task. Because completing different numbers of trials could lead or reflect differential engagement with the task, we will include the effect of the number of blocks completed when modeling performance in the task.

Figure 3B depicts the likelihood of correctly selecting the within-category match in the Inference task across the two lure types and shows that children were more likely to select the within-category match when it was presented in the context of a distant ($M = 0.75$, $SD = 0.22$) than a close ($M = 0.52$, $SD = 0.16$) lure. A model testing the effect of lure distance (close vs. distant) and number of blocks completed on the likelihood of selecting the within-category match showed that lure distance was a

significant predictor of accuracy [$b = 1.13$, $z = 8.34$, $p < 0.0001$, Cohen's $d = 1.22$], but that the number of blocks completed did not significantly predict accuracy in the task [$b = 0.12$, $z = 1.04$, $p = 0.28$]; the model included by-participant random intercepts (the model including random slopes for the effect of lure distance failed to converge). The effect size of the lure distance manipulation was of similar (albeit larger) magnitude relative to when data were collected in person (Fisher et al., 2015).

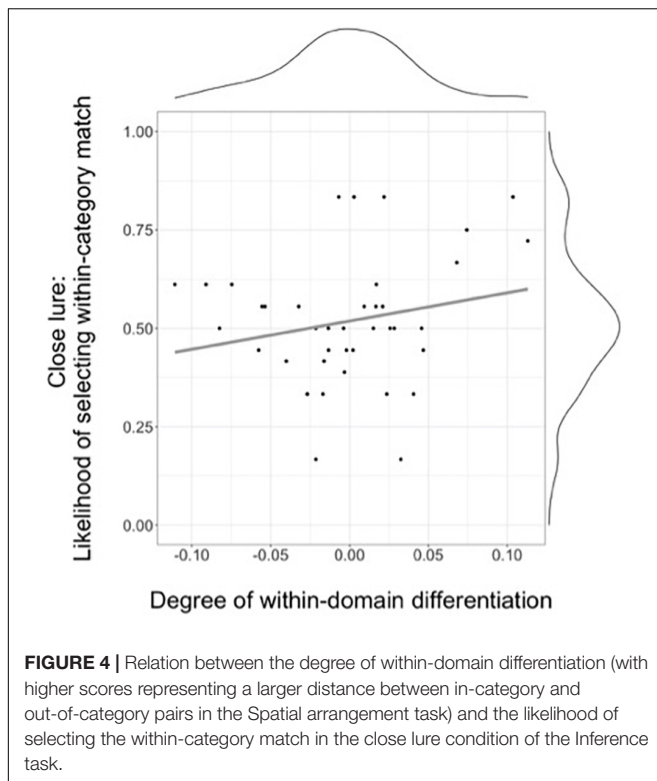
Together, these results provide a conceptual replication of the lure distance effect reported in prior work (Fisher et al., 2015). The comparable results – both conceptually and in magnitude – across means of data collection suggest that remote data collection procedures can be used to examine semantic inferences. These results also suggest that an adaptive procedure in which children decide how many blocks of trials they complete is a viable methodological choice to maximize the number of trials collected while maintaining engagement with the task.

Relation Between Degree of Within-Domain Differentiation and Inferences in the Presence of Close Lures

To examine the relation between a child's within-domain semantic differentiation and the likelihood of inferring that more strongly related items within a domain are more likely to share a property, we calculated a within-domain semantic differentiation score for each child by subtracting the normalized average distance for 'in category' pairs from the normalized average distance for 'out of category' pairs; larger difference scores thus reflect a larger degree of within-domain differentiation. Because the targets in the inference task were all insect 'bugs,' these difference scores only included pairs from the domain of 'bugs.' **Figure 4** shows the association between a child's within-domain differentiation score and their likelihood of selecting the match in the close lure condition, and suggests that there is no such relation. A linear model showed that the within-domain differentiation score was not a significant predictor of a child's average accuracy in the close lure condition [$b = 0.71$, $R^2 = 0.046$, $F(1,38) = 1.23$, $p = 0.19$].

These results suggest that these tasks, as set up for this study, were not able to detect the association between semantic differentiation and semantic inference reported in prior work (Fisher et al., 2015). At first glance this could be taken to indicate that remote data collecting procedures are not well-suited to detect individual differences in semantic structure and/or in semantic inferences. However, it seems more likely that this lack of an association is instead due to methodological choices resulting from the main goals of this study – specifically, the goal of replicating patterns of semantic differentiation in across-versus within-domain distinctions.

As seen in **Figure 4**, the distribution of within-domain difference scores shows a fairly narrow range (-0.11 to 0.11) and is mostly centered around zero – suggesting that most children showed no evidence of differentiating within a domain – making it challenging to examine the role of variability in semantic differentiation. This distribution of scores stands in contrast with



prior work (Fisher et al., 2015), which showed a larger range of differentiation, as well as an association between the two tasks (also see Unger and Fisher, 2019 for related evidence). The observed narrow range and distribution centered at zero is likely due, at least in part, to the fact that children in this age show fairly undifferentiated representations within a domain. However, this weak within-domain differentiation is likely exacerbated by the fact that we tested within- and across-domain differentiation *in the same trial*. We did so because this more closely replicated prior procedures (Vales et al., 2020a,b), but also to decrease the time necessary to complete the spatial arrangement task (and thus decrease possible attrition in the study) – both decisions well-aligned with the goal of replicating previously reported patterns of semantic differentiation. This, however, results in a considerable difference relative to the procedure employed by Fisher et al. (2015), who tested *triads of items* in each trial – thus providing children with a much smaller number of items at a time and thus more degrees of freedom to arrange them. In the case of the spatial arrangement task as designed for this study, the need to attend to both within- and across-domain differentiation, as well as the larger number of cards presented at once, likely reduced the likelihood of detecting individual differences in within-domain differentiation [see Experiment 2 in Vales et al. (2020b) for converging evidence]. Taken together, these results suggest that future work examining semantic structure – and in particular, individual differences in within-domain differentiation in young children – may want to consider whether to assess within-domain differentiation in separate trials and how many items to present in each trial.

DISCUSSION

This manuscript reports a successful conceptual replication of two semantic differentiation effects in 4- to 6-year-old children that were previously reported using in-person data collection. In the spatial arrangement task, children more strongly differentiated across domains relative to within a domain – a pattern of semantic differentiation that replicates prior work (Vales et al., 2020a,b). In the semantic inference task, children's likelihood of selecting a within-domain category match was decreased in the presence of a close (relative to a distant) lure, replicating prior work (Fisher et al., 2015). The conceptual replication of these two effects – which speak to (1) the mechanisms by which organized semantic representations are acquired and (2) the role of organized semantic representations in supporting inferential processes – suggests that such large-sized effects can be successfully reproduced using remote data collection procedures despite the wide variation in the factors that are typically well-controlled during in-person research (such as display size and number of trials) (see also Nussenbaum et al., 2020). These results are also the first evidence that a computerized version of the spatial arrangement method can be successfully completed by children participants, and that an adaptive procedure that allows children to decide how many blocks to complete in the semantic inference task is a promising way to increase the number of trials collected from each participant while maintaining engagement with the task – both important methodological innovations, likely to be useful even in other domains of developmental science.

The use of remote data collection procedures can help strengthen developmental science. By removing a number of barriers to participation, remote data collection has the potential to increase diversity in recruited samples and facilitate the collection of larger sample sizes – both of which are critically necessary. Additionally, as a result of social-distancing measures to mitigate the spread of COVID-19, the field of developmental science is increasing the use of remote data collection procedures. The present results, showing that data collected with children participants remotely is comparable to data obtained from in-person assessments, provide a proof-of-concept that the constructs measured by these tasks can be successfully assessed remotely and thus increase the confidence that developmental scientists can continue to accumulate and integrate knowledge across different mediums of data collection.

It is important to note that the effects we set out to replicate were medium-sized; future work should evaluate if smaller-sized effects can also be replicated under the more variable testing conditions inherent to remote testing. Similarly, other tasks might be more sensitive to these more variable testing conditions; for example, increased distractions in the home environment might be more problematic in the context of experimental tasks requiring the collection of reaction time (but see Nussenbaum et al., 2020). Future work should consider these possible limiting factors when planning online data collection. We also note that not all children completed both tasks, with about 20% of children who completed the spatial arrangement task not completing the inference task. Prior work examining the relation

between these two tasks (Fisher et al., 2015) conducted the two tasks in two separate sessions, as the study included numerous measures at multiple time points. As such, we do not know whether the attrition rate observed here would be similar to in-person data collection procedures. Future work intending to collect multiple measures per participant within the same study session should consider the attrition rate observed here and decide whether conducting multiple sessions may be a better approach to their goals.

Remote data collection procedures by themselves will not be sufficient to realize the promise of increasing diversity in study samples. The sample in this study was a convenience sample resulting from an ongoing partnership with the science outreach team at a local botanical garden, and thus we did not aim to obtain a geographically diverse sample (although some families joined from out-of-state, which would have been unlikely had the programs taken place in person). When planning this collaboration, we took steps to increase diversity in the demographics of children participants, both through publicizing the camps in underserved neighborhoods and by reducing enrollment costs – and these efforts seem to have been successful to some extent, as we saw an increase in non-white participants relative to a prior collaboration (Vales et al., 2020a) and considerable variability in the neighborhoods (i.e., zip codes) where the participants lived. However, because these camps were moved to a remote medium as a result of social-distancing guidelines due to the COVID-19 pandemic in the Spring and Summer of 2020, there were considerable changes in enrollment as family and childcare circumstances quickly changed. This makes it difficult to know whether our efforts to broaden participation could have been more successful under different circumstances. Indeed, as Lourenco and Tasimi (2020) note, researchers must continue to take steps to ensure equitable access for families from disadvantaged backgrounds, especially during a pandemic when access to internet might be even more challenging (e.g., libraries might not be open to the public).

The current study failed to find an association between the degree of a child's within-domain differentiation and their likelihood of selecting the matching within-domain item in the presence of a close (i.e., belonging to the same-domain) lure. Although this could be taken to indicate that remote data collection procedures are not well-suited to detect individual differences in these two processes, it seems more likely that the lack of an association between the two tasks is instead due to the limited range of scores and a distribution centered around zero that was observed for the within-domain difference scores. We believe these undifferentiated scores are a result of both weak within-domain differentiation (consistent with the patterns found in the spatial arrangement task) and the fact that both within- and across-domain differentiation were tested *in the same trial*, which reduced the degrees of freedom for arranging individual cards. This is a crucial difference relative to prior work (Fisher et al., 2015), and in requiring children to simultaneously attend to both distinctions might have reduced the odds that children noticed within domain distinctions. Prior work using this task suggests that these are important methodological considerations (Vales et al., 2020b), and we believe future work intending to use remote assessments of semantic structure and

semantic inferences should consider the goals of the assessments when deciding whether to examine within- and across-domain differentiation in the same or separate trials.

The recruitment strategy we used – recruiting children participating in a virtual enrichment program – can also be a useful tool for researchers to maintain and extend their partnerships with community settings during the current limitations to in-person testing. Over the course of only three weeks, with a single 2.5 h-long session involving 5–7 researchers each week, we recruited and tested more than 50 children. The researcher involvement was fairly minimal, and it is likely that with some improvements to the usability of the tasks it would be possible for children to complete these tasks without any researcher involvement. Partnerships between basic science researchers and educators are an important component of developmental science and can be mutually beneficial for the researchers and the educators (Osberg, 1998; Callanan, 2012; Haden, 2020; Mulvey et al., 2020). The COVID-19 pandemic has propelled the development of virtual learning programs (Bell, 2020); this study illustrates how researchers can leverage this reality to maintain existing partnerships within their local communities and possibly develop new ones with science centers that were previously geographically inaccessible – and in so doing, study developmental change in ecologically valid settings (Golinkoff et al., 2017).

In sum, the current results suggest that the spatial arrangement task and the semantic inference task can be successfully employed to remotely assess semantic structure. This allows future work using these tasks to be aggregated with prior work using in-person data collection procedures. This also provides researchers with alternative ways to recruit larger and more diverse samples, and thus continue to strengthen practices in developmental science.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://osf.io/67gtc/>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Carnegie Mellon University Institutional Review Board. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

CV, CW, and AF designed the study. CV and CW processed the data and performed the statistical analyses. CV wrote the first version of the manuscript. CW wrote sections of the manuscript. All the authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was supported by the National Science Foundation (award 1918259 to AF and CV) and by the James S. McDonnell Foundation (Scholar Award 220020401 to AF).

ACKNOWLEDGMENTS

We thank Kaitlynn Cooper, Juan Forero, Emma Gurchiek, Jasmine Liu, Xiaoying Meng, Suanna Moron, Oceann Stanley, and Kate Zhao for their help with stimuli creation, task

development, and data collection; the Phipps Conservatory and Botanical Gardens staff for facilitating participant recruitment and testing; and the children and families who participated in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.697550/full#supplementary-material>

REFERENCES

- Alibali, M. W., and Nathan, M. J. (2010). Conducting research in schools: a practical guide. *J. Cogn. Dev.* 11, 397–407. doi: 10.1080/15248372.2010.516417
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Memory Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Statist. Softw.* 67, 1–48.
- Bell, J. (2020). *National Science Foundation-Funded Projects With Online Learning Products & Resources. Informal Science*. Available online at: <https://www.informalscience.org/news-views/nsf-aisl-funded-projects-online-learning-products-resources> (accessed March 26, 2021).
- Belsky, J. (1980). Mother-infant interaction at home and in the laboratory: a comparative study. *J. Genet. Psychol.* 137, 37–47. doi: 10.1080/00221325.1980.10532800
- Bjorklund, D. F., and Jacobs, J. W. (1985). Associative and categorical processes in children's memory: the role of automaticity in the development of organization in free recall. *J. Exp. Child Psychol.* 39, 599–617. doi: 10.1016/0022-0965(85)90059-1
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* 14, 365–376. doi: 10.1038/nrn3475
- Callanan, M. A. (2012). Conducting cognitive developmental research in museums: theoretical issues and practical considerations. *J. Cogn. Dev.* 13, 137–151. doi: 10.1080/15248372.2012.666730
- Cates, C. B., Weisleder, A., Johnson, S. B., Seery, A. M., Canfield, C. F., Huberman, H., et al. (2018). Enhancing parent talk, reading, and play in primary care: sustained impacts of the video interaction project. *J. Pediatr.* 199, 49–56. doi: 10.1016/j.jpeds.2018.03.002
- Chen, W., and Adler, J. L. (2019). Assessment of screen exposure in young children, 1997 to 2014. *JAMA Pediatr.* 173, 391–393. doi: 10.1001/jamapediatrics.2018.5546
- Chuey, A., Lockhart, K., Sheskin, M., and Keil, F. (2020). Children and adults selectively generalize mechanistic knowledge. *Cognition* 199:104231. doi: 10.1016/j.cognition.2020.104231
- Clark, E. V. (1973). "What's in a word? on the child's acquisition of semantics in his first language," in *Cognitive Development and the Acquisition Of Language*, ed. T. E. Moore (New York, NY: Academic Press), 65–110. doi: 10.1016/b978-0-12-505850-6.50009-8
- Coley, J. D. (2012). Where the wild things are: informal experience and ecological reasoning. *Child Dev.* 83, 992–1006. doi: 10.1111/j.1467-8624.2012.01751.x
- DeBolt, M. C., Rhemtulla, M., and Oakes, L. M. (2020). Robust data and power in infant research: a case study of the effect of number of infants and number of trials in visual preference procedures. *Infancy* 25, 393–419. doi: 10.1111/inf.12337
- Fernald, A. (2010). Getting beyond the "convenience sample" in research on early cognitive development. *Behav. Brain Sci.* 33:91. doi: 10.1017/s0140525x10000294
- Fisher, A., Thiessen, E., Godwin, K., Kloos, H., and Dickerson, J. (2013). Assessing selective sustained attention in 3- to 5-year-old children: evidence from a new paradigm. *J. Exp. Child Psychol.* 114, 275–294. doi: 10.1016/j.jecp.2012.07.006
- Fisher, A. V., Godwin, K. E., and Matlen, B. J. (2015). Development of inductive generalization with familiar categories. *Psychonomic Bull. Rev.* 22, 1149–1173. doi: 10.3758/s13423-015-0816-5
- Forrester, S. E. (2015). Selecting the number of trials in experimental biomechanics studies. *Int. Biomechan.* 2, 62–72. doi: 10.1080/23335432.2015.1049296
- Friend, M., and Keplinger, M. (2003). An infant-based assessment of early lexicon acquisition. *Behav. Res. Methods Instruments Comp.* 35, 302–309. doi: 10.3758/bf03202556
- Gelman, S. A., and Markman, E. (1986). Categories and induction in young children. *Cognition* 23, 183–209. doi: 10.1016/0010-0277(86)90034-x
- Gershon, R. C., Cella, D., Fox, N. A., Havlik, R. J., Hendrie, H. C., and Wagster, M. V. (2010). Assessment of neurological and behavioural function: the NIH Toolbox. *Lancet Neurol.* 9, 138–139. doi: 10.1016/s1474-4422(09)70335-7
- Gesell, A. (1932). How science studies the child. *Sci. Monthly* 34, 265–267.
- Gibbons, R. D., Hedeker, D. R., and Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *J. Educ. Statist.* 18, 271–279. doi: 10.2307/1165136
- Gobbo, C., and Chi, M. (1986). How knowledge is structured and used by expert and novice children. *Cogn. Dev.* 1, 221–237. doi: 10.1016/s0885-2014(86)80002-8
- Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behav. Res. Methods Instruments Comp.* 26, 381–386. doi: 10.3758/bf03204653
- Golinkoff, R. M., Hirsh-Pasek, K., Grob, R., and Schlesinger, M. (2017). Oh, the places you'll go" by bringing developmental science into the world! *Child Dev.* 88, 1403–1408. doi: 10.1111/cdev.12929
- Haden, C. A. (2020). *Developmental Science Research with Children's Museums, not just at Them [Peer commentary on the article "Exploration, Explanation, and Parent-Child Interaction in Museums" by MA Callanan, CH Legare, DM Sobel, G. Jaeger, S. Letourneau, SR McHugh, A. Willard, et al.]*. Washington, DC: SRCD.
- Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., and Hilbig, B. E. (2019). *Lab.js: A Free, Open, Online Study Builder*. Available online at: <https://doi.org/10.31234/osf.io/fqr49>
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83. doi: 10.1017/s0140525x0999152x
- Hills, T. T., Maouene, M., Maouene, J., Sheya, A., and Smith, L. (2009). Categorical structure among shared features in networks of early-learned nouns. *Cognition* 112, 381–396. doi: 10.1016/j.cognition.2009.06.002
- Jenkins, G. W., Samuelson, L. K., Smith, J. R., and Spencer, J. P. (2015). Non-Bayesian noun generalization in 3- to 5-year-old children: probing the role of prior knowledge in the suspicious coincidence effect. *Cogn. Sci.* 39, 268–306. doi: 10.1111/cogs.12135
- Kemp, C., and Tenenbaum, J. B. (2008). The discovery of structural form. *Proc. Natl. Acad. Sci. U.S.A.* 105, 10687–10692.
- Klein, R. P., and Dufree, J. T. (1979). Comparison of attachment behaviors in home and laboratory. *Psychol. Rep.* 44, 1059–1064. doi: 10.2466/pr0.1979.44.3c.1059
- Koch, A., Speckmann, F., and Unkelbach, C. (2020). *Q-SpAM: How to efficiently Measure Similarity in Online Research*. Thousand Oaks, CA: Sage Publications Inc.
- Leshin, R., Leslie, S. J., and Rhodes, M. (2021). Does it matter how we speak about social kinds? a large, pre-registered, online experimental study of how language shapes the development of essentialist beliefs. *Child Dev.* Online ahead of print

- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Mulvey, K. L., McGuire, L., Hoffman, A. J., Hartstone-Rose, A., Winterbottom, M., Balkwill, F., et al. (2020). Learning hand in hand: engaging in research–practice partnerships to advance developmental science. *New Direct. Dhild Adolescent Dev.* 2020, 125–134. doi: 10.1002/cad.20364
- Nielsen, M., Haun, D., Kärtner, J., and Legare, C. H. (2017). The persistent sampling bias in developmental psychology: a call to action. *J. Exp. Child Psychol.* 162, 31–38. doi: 10.1016/j.jecp.2017.04.017
- Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., and Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychol.* 6:17213.
- Osberg, S. (1998). Shared lessons and self-discoveries: what research has taught Children's Discovery Museum. *J. Museum Educ.* 23, 19–20. doi: 10.1080/10598650.1998.11510367
- Ossmer, C. (2020). Normal development: the photographic dome and the children of the yale psycho-clinic. *Isis* 111, 515–541. doi: 10.1086/711127
- Perry, L. K., Samuelson, L. K., and Burdinie, J. B. (2014). Highchair philosophers: the impact of seating context-dependent exploration on children's naming biases. *Dev. Sci.* 17, 757–765. doi: 10.1111/desc.12147
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Rideout, V. (2017). *The Common Sense Census: Media use by Kids Age Zero to Eight*. San Francisco, CA: Common Sense Media, 263–283.
- Rogers, T. T., and McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press.
- Santolin, C., Garcia-Castro, G., Zettersten, M., Sebastian-Galles, N., and Saffran, J. R. (2021). Experience with research paradigms relates to infants' direction of preference. *Infancy* 26, 39–46. doi: 10.1111/inf.12372
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/opmi_a_00002
- Sheskin, M., and Keil, F. (2018). TheChildLab.com: a video chat platform for developmental research. *PsyArXiv [preprint]* doi: 10.31234/osf.io/rn7w5
- Sugden, N. A., and Moulson, M. C. (2015). Recruitment strategies should not be randomly selected: empirically improving recruitment success and diversity in developmental psychology research. *Front. Psychol.* 6:523. doi: 10.3389/fpsyg.2015.00523
- Torchiano, M. (2020). *effsize: Efficient Effect Size Computation. R package*.
- Tversky, B. (1985). Development of taxonomic organization of named and pictured categories. *Dev. Psychol.* 21, 1111–1119. doi: 10.1037/0012-1649.21.6.1111
- Unger, L., and Fisher, A. V. (2019). Rapid, experience-related changes in the organization of children's semantic knowledge. *J. Exp. Child Psychol.* 179, 1–22. doi: 10.1016/j.jecp.2018.10.007
- Vales, C., States, S. L., and Fisher, A. V. (2020a). Experience-driven semantic differentiation: effects of a naturalistic experience on within- and across-domain differentiation in children. *Child Dev.* 91, 733–742. doi: 10.1111/cdev.13369
- Vales, C., Stevens, P., and Fisher, A. V. (2020b). Lumping and Splitting: developmental changes in the structure of children's semantic networks. *J. Exp. Child Psychol.* 199:104914. doi: 10.1016/j.jecp.2020.104914
- Varga, D. (2011). Look–normal: the colonized child of developmental science. *History Psychol.* 14:137. doi: 10.1037/a0021775
- Waxman, S. R., and Namy, L. L. (1997). Challenging the notion of a thematic preference in young children. *Dev. Psychol.* 33:555. doi: 10.1037/0012-1649.33.3.555

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Vales, Wu, Torrance, Shannon, States and Fisher. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Unsupervised Online Assessment of Visual Working Memory in 4- to 10-Year-Old Children: Array Size Influences Capacity Estimates and Task Performance

Shannon Ross-Sheehy^{1*}, Esther Reynolds¹ and Bret Eschman²

¹ Department of Psychology, University of Tennessee, Knoxville, Knoxville, TN, United States, ² Department of Psychology, Florida International University, Miami, FL, United States

OPEN ACCESS

Edited by:

Natasha Kirkham,
Birkbeck, University of London,
United Kingdom

Reviewed by:

Chen Cheng,
Boston University, United States
Christian H. Poth,
Bielefeld University, Germany

*Correspondence:

Shannon Ross-Sheehy
rosssheehy@utk.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 07 April 2021

Accepted: 26 May 2021

Published: 06 August 2021

Citation:

Ross-Sheehy S, Reynolds E and
Eschman B (2021) Unsupervised
Online Assessment of Visual Working
Memory in 4- to 10-Year-Old Children:
Array Size Influences Capacity
Estimates and Task Performance.
Front. Psychol. 12:692228.
doi: 10.3389/fpsyg.2021.692228

The events of the COVID-19 Pandemic forced many psychologists to abandon lab-based approaches and embrace online experimental techniques. Although lab-based testing will always be the gold standard of experimental precision, several protocols have evolved to enable *supervised* online testing for paradigms that require direct observation and/or interaction with participants. However, many tasks can be completed online in an *unsupervised* way, reducing reliance on lab-based resources (e.g., personnel and equipment), increasing flexibility for families, and reducing participant anxiety and/or demand characteristics. The current project demonstrates the feasibility and utility of unsupervised online testing by incorporating a classic change-detection task that has been well-validated in previous lab-based research. In addition to serving as proof-of-concept, our results demonstrate that large online samples are quick and easy to acquire, facilitating novel research questions and speeding the dissemination of results. To accomplish this, we assessed visual working memory (VWM) in 4- to 10-year-old children in an unsupervised online change-detection task using arrays of 1–4 colored circles. Maximum capacity (max K) was calculated across the four array sizes for each child, and estimates were found to be on-par with previously published lab-based findings. Importantly, capacity estimates varied markedly across array size, with estimates derived from larger arrays systematically underestimating VWM capacity for our youngest participants. A linear mixed effect analysis (LME) confirmed this observation, revealing significant quadratic trends for 4- through 7-year-old children, with capacity estimates that initially increased with increasing array size and subsequently decreased, often resulting in estimates that were *lower* than those obtained from smaller arrays. Follow-up analyses demonstrated that these regressions may have been based on explicit guessing strategies for array sizes perceived too difficult to attempt for our youngest children. This suggests important interactions between VWM performance, age, and array size, and further suggests estimates such as *optimal array size* might capture both *quantitative* aspects of VWM performance and *qualitative* effects of

attentional engagement/disengagement. Overall, findings suggest that unsupervised online testing of VWM produces reasonably good estimates and may afford many benefits over traditional lab-based testing, though efforts must be made to ensure task comprehension and compliance.

Keywords: visual working memory, child development, online assessment, cognitive development, capacity estimates

INTRODUCTION

Infant research is difficult for many reasons. Access to public records is increasingly restricted, contact information is often unpublished, and in many areas, families and communities are becoming wary of privacy concerns and university sponsored research. In addition, the reality of dual-income families continues to make lab-based testing in the early months and years of life a logistical challenge. Although the gold standard of experimental precision will likely always center around lab-based techniques, changing work and family dynamics necessitates a re-evaluation of the gold-standard approach.

The events of the COVID-19 Pandemic forced many psychologists to abandon lab-based techniques and embrace online experimental approaches. This has been particularly difficult for developmentalists, as many infant and child-based testing techniques rely on looking time or eye-tracking methodologies. Fortunately, many innovative approaches have been developed that allow for live face-to-face testing (i.e., *supervised* testing), including commercial video conferencing options like *Zoom* and *Microsoft Teams*, and homegrown software solutions such as *Lookit* (<https://lookit.mit.edu>). While these approaches facilitate remote observation of the child engaging in the task, they involve many of the same resources as lab-based work, including dedicated experimenters and observers to run test sessions, and pre-scheduled appointments with families. However, for tasks that can be adapted to rely solely on behavioral responses (key presses, mouse clicks, touch screens, etc.), it is possible to do remote online testing in an *unsupervised* way. We report here results from a large-scale unsupervised online change-detection task assessing visual working memory (VWM) development continuously from 4 to 10 years of age.

There are several practical benefits of conducting unsupervised online research. First, it increases session flexibility, allowing participation at optimal times such as after naps, on a rainy Saturday afternoon, or when network traffic is low. Second, it allows for home-based testing, which in addition to being more convenient for parents and children, may decrease the anxiety and demand characteristics that are inevitably a part of supervised testing procedures. Third, unsupervised at-home testing may allow participation from a wider range of children, both neuro-typical and neuro-atypical, and allows for rapid testing over a broad range of ages.

In addition to these practical advantages, there are a host of scientific benefits that may increase data validity and facilitate novel research questions. For example, this approach reduces

the time and resources necessary to acquire large sample sizes, increasing power and replicability for even relatively small effects. This speeds dissemination of research findings, and may facilitate novel findings and theory building. Unsupervised online testing can also be conducted regionally, nationally, or even internationally without regard to time zone constraints. In addition to facilitating epidemiological approaches to the study of development, online testing can improve racial, ethnic, and socioeconomic diversity, something that is profoundly lacking from most lab-based research samples. Although access to computers and internet connections may vary across these diverse populations, it is possible for participants to conduct these tasks using a mobile device or tablet, a friend or family member's computer, or public resources such as school, library or community computer banks. Finally, online testing allows the explicit testing of environment factors such as screen size, stimulus size, and method of response (e.g., mouse, keyboard or touchscreen). These features are often either ignored completely or held constant in lab-based tasks, despite the fact that changes in these simple task features might critically influence performance. This form of apparatus diversity additionally ensures that findings are robust, and context independent.

There are of course some drawbacks to unsupervised online testing, including lack of control (Anwyl-Irvine et al., 2020a) and the possibility of parental interference and/or non-compliance with experimental procedures. All of these can be ameliorated to some extent using tools present in most modern online experimental testing suites (e.g., *Gorilla.sc* and *LabVanced.com*), including ability to collect webcam video and to “calibrate” or scale the stimuli based on the estimated screen size. It is also possible to use *indirect* measures to identify questionable data, such as participants whose response times are either too fast to be plausibly completed by the participant (i.e., parental interference), or to reflect effortful decision and response selection (i.e., random button presses). We incorporated several of these approaches in the current project. However, one of the most challenging and underappreciated aspects of successful online testing, is accurately conveying task instructions to the children and to the parents who function as *ad hoc* experimenters. In contrast to supervised testing approaches, it is impossible to gauge understanding and solicit questions from families during unsupervised testing. Thus, it is critically important that the task be piloted in the lab with the target age demographic, to reveal confusing and problematic aspects of the task instructions. This process also facilitates the development of videos and practice trials that maximally enhance task understanding.

Choice of task is also a key factor. The current project incorporates an unsupervised online testing approach to assess development of VWM, which is quite easily adapted to rely solely on behavioral responses (mouse or keyboard clicks, or touches). This task was chosen, because VWM is an essential visuospatial ability that shows substantial development over the first several years of life (Ross-Sheehy et al., 2003; Gathercole et al., 2004; Oakes et al., 2006; Simmering and Spencer, 2008; Simmering and Perone, 2013; Buss et al., 2018; Ross-Sheehy and Eschman, 2019; Reyes et al., 2020), and developmental profiles have already been established across a range of ages (e.g., Cowan et al., 2005; Simmering, 2012). VWM is an active form of short-term memory, that supports the processing of visual spatial information in service of a task or goal (Luck and Vogel, 1997). Many tasks that support early learning rely heavily on VWM, including visual comparison, categorization, spatial navigation, visual search, object learning, spatial reasoning, and math. Thus, VWM is a critically important component of general cognitive development.

Much research has tied VWM to later academic achievement. For example, Bull (2008) found that VWM performance in preschool predicted math problem solving at 8 years of age. Similarly, others have found that VWM in 7- to 14-year-olds predicted performance on a national curriculum math test (Jarvis and Gathercole, 2003). These basic findings have now been replicated numerous times, with most results demonstrating an important connection between early VWM and later math achievement (Tsubomi and Watanabe, 2017; Giofrè et al., 2018; Allen et al., 2019; Chan and Wong, 2019; Kytälä et al., 2019; Carr et al., 2020). VWM in adults is related to measures of fluid intelligence (Fukuda et al., 2010), and the development of VWM is distinct from verbal WM (Gathercole and Baddeley, 1993; Jarvis and Gathercole, 2003; Giofrè et al., 2018; Kytälä et al., 2019) and executive function aspects of WM (Jarvis and Gathercole, 2003; Gathercole et al., 2004). Thus, early and frequent access to online VWM assessment tools could significantly enhance detection and possibly intervention for children at risk of cognitive delay. Although the literature on WM training interventions is mixed, recent ERP work with adults demonstrates hopeful evidence of persistent VWM training benefits (Zhang et al., 2020).

The Current Project

The goal of the current project is to demonstrate the feasibility and validity of unsupervised online testing approaches in child populations, by incorporating a canonical lab-based change-detection task previously used in infant, child and adult populations (Luck and Vogel, 1997; Cowan et al., 2005; Riggs et al., 2006; Ross-Sheehy and Eschman, 2019). The task was adapted for online testing and was used to assess VWM development from 4 to 10 years of age. Our task incorporated a whole-report change-detection approach, meaning all array items were present both in the sample and test arrays, and the child's job was to determine if anything changed from the sample to the test array. Although many adult change-detection tasks now utilize a single-probe or partial report approach (Rouder et al., 2011), we opted to incorporate the whole-report

approach for two reasons: First, pilot studies conducted in our lab suggested that younger children struggled to understand the concurrence between sample and test arrays, and altering test arrays might further disrupt within-trial continuity for our youngest participants. Second, this task has already been used successfully in both infant and adult participants (Ross-Sheehy and Eschman, 2019), facilitating the examination of capacity development from infancy to childhood and beyond.

METHODS

Participants

Our participant pool was a sample of convenience and included all families of children born in local or neighboring counties who had previously expressed an interest in study participation. All registered families with children between the ages of 4 and 11 years during our 6-month data collection window were contacted via email and invited to participate. Of the 2,949 families contacted, 9.93% agreed to participate, resulting in a sample of 297 children (see **Table 1** for demographics). Unlike standard lab tasks, data quality could not be assessed until after participation was complete. As a first step, we examined survey responses for each participant. This resulted in the exclusion of children due to frustration or inability to understand the task ($n = 3$), diagnosis of developmental delay ($n = 1$) or autism spectrum disorder ($n = 5$), incorrect age ($n = 1$), or completing the task using a mobile phone ($n = 1$). We next assessed general task performance by examining the number trials completed out of 80 possible trials, as well as general performance (hit, miss, correct rejection, and false alarm rates). We excluded children who did not complete at least 3 blocks of trials ($n = 18$, $M_{\text{trials}} = 13$, $SD_{\text{trials}} = 3.7$), and children who selected only a single response button ($n = 1$). Although several children reported a family history of colorblindness ($n = 11$) an examination of their results revealed typical patterns of responding, so they were retained in the sample. Task engagement for the final sample of 267 subjects was very high, $M_{\text{trials}} = 76.67$, $SD_{\text{trials}} = 12.81$.

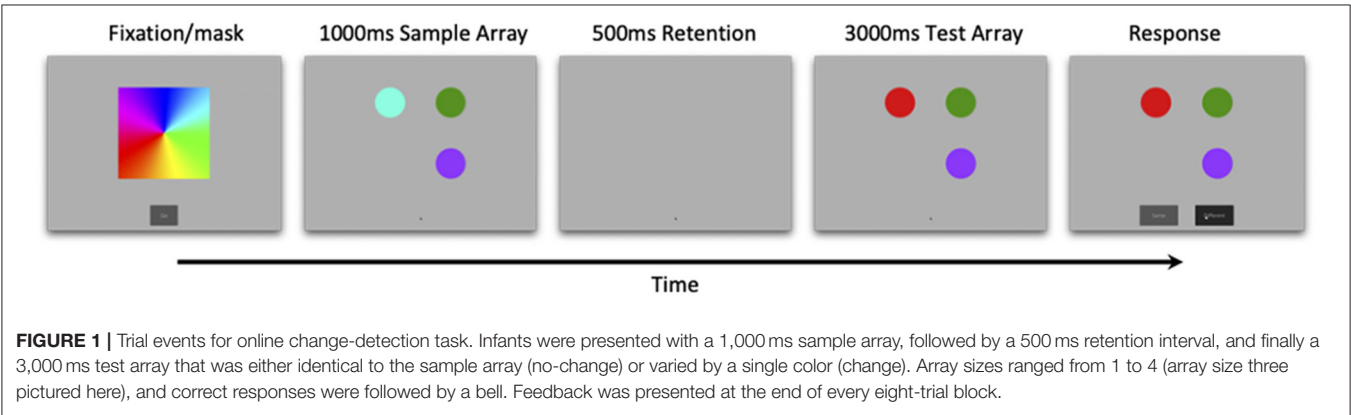
Stimuli

Stimuli for this study were based on Ross-Sheehy and Eschman (2019). Each trial started with a colorful spinning pinwheel that oriented attention, and served as a between-trial mask. Participants were then tested in a change-detection paradigm consisting of a 1,000 ms sample array containing 1–4 colored circles, followed by a 500 ms retention interval, and finally a 3,000 ms test array that was either identical to the sample array (no-change trials) or included a color change presented at a random location (change trials). After 3,000 ms two response buttons appeared underneath the test array, labeled “same” or “different” (**Figure 1**). Participants saw up to 10 blocks of trials and each block consisted of one of every possible trial type (array size 1, 2, 3, 4, change and no-change) presented randomly.

The circles in both sample and test arrays were presented at 45°, 135°, 225°, and 315° relative to the center of the display, but were constrained to stay within the boundary of the colorful pinwheel perceptual mask. Circles consisted of eight highly discriminable colors (blue, orange, red, yellow, purple,

TABLE 1 | Participant counts and demographics by age (years).

| Age | N | M | SD | Min | Max | SES (annual) | | | Race | | | | | Ethnicity | | |
|-----|----|-------|------|-------|-------|--------------|-------|-------|-------|------------|-------|---------------|-------|------------|----|----------|
| | | | | | | Female | <80 K | ≥80 K | Asian | Am. Indian | Black | Pac. Islander | White | Mult. Race | NA | Hispanic |
| 4 | 43 | 4.59 | 0.25 | 4.02 | 4.98 | 42% | 23% | 77% | 9% | 2% | 2% | 2% | 81% | 0% | 2% | 7% |
| 5 | 56 | 5.50 | 0.32 | 5.01 | 5.99 | 53% | 20% | 80% | 7% | 2% | 5% | 0% | 86% | 0% | 0% | 5% |
| 6 | 50 | 6.42 | 0.28 | 6.00 | 6.98 | 48% | 33% | 67% | 4% | 0% | 2% | 0% | 90% | 2% | 2% | 4% |
| 7 | 31 | 7.54 | 0.25 | 7.02 | 7.98 | 38% | 24% | 76% | 9% | 3% | 13% | 0% | 72% | 0% | 3% | 0% |
| 8 | 32 | 8.43 | 0.27 | 8.00 | 8.97 | 38% | 17% | 83% | 6% | 6% | 3% | 0% | 78% | 3% | 3% | 0% |
| 9 | 28 | 9.45 | 0.32 | 9.00 | 9.99 | 70% | 26% | 74% | 3% | 0% | 13% | 0% | 83% | 0% | 0% | 3% |
| 10 | 26 | 10.36 | 0.30 | 10.03 | 10.99 | 52% | 9% | 91% | 4% | 0% | 11% | 4% | 81% | 0% | 0% | 7% |



cyan, green and magenta) and were presented against a gray background. Circle locations and colors were chosen randomly without replacement for each trial using a custom python script, and circles for array size 2 were constrained to contiguous locations only (no obliques). Although Gorilla.sc does allow for active stimulus scaling based on visual angle, this scaling operates on individual display objects (i.e., individual circles in our case) and does not address the relative spacing between objects. That is, even though the individual circles might successfully be scaled based on visual angle, the gaps between them were not. Given chunking efficacy might vary with relative circle proximity we chose not to incorporate object-based scaling, and instead opted for passive scaling of the entire configuration based on monitor size. Although this did not explicitly equate visual angle across participants, participants with smaller screens (e.g., laptops or iPads) generally sat closer to the screen, roughly equating visual angle and preserving the relative spaces between the circles.

Engaging sounds were presented during both the sample and test arrays to increase interest in the task, highlight cohesion and alignability between sample and test arrays, and to emphasize the change detection judgment during test array. The sample array sound was an ascending slide whistle that continued through both the sample and gap intervals, followed immediately by a “bloop” sound simultaneous with the onset of the test array. A reward tone immediately followed a correct response, and consisted of a pleasant 630 ms bell tone with a frequency of ~2,300 Hz. There was no feedback given for incorrect trials.

Procedure

All methods and procedures were approved by University of Tennessee IRB #17-03545. Parents were invited to participate based on previous participation in one of the University of Tennessee Child Development Research Labs. Parents of children 4–10 years were sent an email inviting them to participate in an at-home test of cognitive development. If interested, parents clicked a link, and were taken immediately to an online consent form (children aged 7 and older were additionally assented). Upon completion of the consent, parents filled out a general demographic questionnaire, and were then routed to the online experiment portal (*Gorilla.sc*; Anwyl-Irvine et al., 2020b). Parents and children were given general instructions regarding the online browser-based “memory game,” and informed that the game could be quit and resumed if the child became bored, or if network congestion was high. Parents were then presented with several “get ready” screens, instructing them to ensure their child had a distraction free environment, that their browser was in full screen mode, and that their computer’s sound was set at an appropriate level. Prior to online testing, pilot testing occurred in the lab with 3- and 4-year-old children, parents, and adult participants. These experiences helped us determine the youngest feasible age for unsupervised testing, and informed the video demonstration and instructions that appeared prior to the onset of the task. Previous online testing experience suggested this process to be critically important in preventing frustration and enhancing understanding of the task expectations. Parents and

children were then presented with a video demonstration of the memory game:

“This colorful pinwheel will appear at the beginning of each trial. Press “Go” to begin. [child presented with dynamic image of spinning pinwheel and “go” button]. For each trial, some circles will briefly appear [child is shown a sample array containing colored circles], then disappear [child is shown blank display], then reappear [child is shown test array identical to the sample array with the exception of a single color change. After a brief delay, two response buttons were presented underneath the test array, one labeled “Same” one labeled “Different”]. Your child’s job is to determine if the circles stayed the same, or if one of them changed color. Have your child respond aloud, then click “Same” or “Different” to indicate their response. If your child is correct, a bell will ring [animation of mouse cursor clicking the “Different” button, followed by a bell]. The circles blink quickly, so be sure not to start the trial until your child is ready! We will vary the position of the circles, and how many appear [children and parents shown several additional demonstration trials]. Remember, this was designed to be challenging! If your child is unsure, encourage them to guess.”

Parents and children could watch the video as many times as necessary before proceeding to the practice trials. Practice trials were identical to task trials, however additional instructions were included at the top of each display. Parents clicked “Continue” when their child was ready to begin the task trials. To keep engagement high, children were presented with a performance screen after the completion of each block. This screen provided encouraging feedback, a progress bar, and the child’s accuracy. It also included two buttons, one to continue the task trials, and one to end the task early. Parents were instructed to end the trials early if their child became uninterested, or no longer wished to participate. The task took an average of 9.74 min to complete ($SD = 2.9$).

Immediately after task completion, parents and participants were administered a brief survey that included a comment field and two questions assessing enjoyment and comprehension (5-point Likert scale, with one representing least possible enjoyment/understanding, and 5 representing greatest possible enjoyment/understanding). Average ratings for enjoyment ($M = 3.6$, $SD = 1.17$) and task comprehension ($M = 4.39$, $SD = 1.02$) suggested parents and children understood the task, and enjoyed it to a reasonable extent. After participation, parents were emailed a \$10 Amazon.com gift card to share with their child.

Two split-half reliability estimates were computed using mean proportion correct at each set size. The first analysis compared accuracy across even and odd trials (i.e., internal consistency) and the second compared accuracy across the first and last half of the trials (i.e., time effects). Cronbach’s alpha indicated good internal consistency between even and odd trials, $\alpha = 0.712$, good reliability over time, $\alpha = 0.730$. Although mean proportion correct was slightly higher for the first half of the experiment ($M = 0.888$, $SD = 0.145$) compared to the last half of the experiment ($M = 0.882$, $SD = 0.147$), this difference was not significant, $t_{(1,059)} = 1.465$, $p = 0.143$.

RESULTS

Raw response times were examined prior to analysis. This revealed one 8-year-old outlier with implausibly high performance (mean response time = 155 ms, perfect performance across all 4 array sizes), who was subsequently removed from our analysis. All other responses conformed to typical developmental patterns (**Figure 2**). We estimated VWM capacity (k) using Pashler’s equation (Pashler, 1988) with $k = N \times (H - FA) / (1 - FA)$, where N = array size, H = hit rate (proportion of change trials in which color change was correctly detected), and FA = false-alarm rate (proportion of no change trials in which color change was erroneously detected). We calculated maximum capacity for each child ($max K$) as the highest capacity estimate produced across all four array sizes. Although there is considerable debate regarding the discrete slots assumptions of Pashler’s approach (Cowan, 2001; Bays and Husain, 2008; Zhang and Luck, 2008; Rouder et al., 2011), this equation is convenient as it incorporates multiple sources of information and is easier to interpret than accuracy or sensitivity measures such as A' or d' . However, Pashler’s equation does not penalize false alarm rates in cases where hit rates were very high. This is one reason why Pashler’s equation may slightly overestimate capacity, particularly in child samples. For this reason, it is important to prescreen results and identify any participants who may have chosen the same response for every trial. This may also help identify children who were confused by the task.

Assessing Data Quality, Task Validity, and Environment Variables

Does Unsupervised Testing Produce Plausible VWM Capacity Estimates?

Because this was an unsupervised task, it was important to assess task performance and compliance, as well as general capacity estimates. A Pearson bivariate correlation revealed a moderate correlation between age and trial counts, $r = 0.230$, $p < 0.001$, with younger children completing fewer trials than older children (**Table 2**). Although 90.4% of participants completed all 80 trials, the 26 participants who completed fewer than 80 trials were relatively young, $M_{age} = 5.68$, $SD_{age} = 1.15$. In addition, younger children took longer to respond on average than older children, $r = -0.565$, $p < 0.001$. This finding is not unique to online testing paradigms, and suggests that relatively slow responses may have contributed to increased task fatigue for the youngest children. Importantly, results for maximum capacity ($max K$) revealed a strong positive correlation with age (**Figure 3**). These estimates are consistent with previously published findings for children of this age, validating this general approach (Simmering, 2012, 2016; Buss et al., 2018).

Do Screen Size and Response Mode Influence VWM Capacity Estimates?

One of the drawbacks of at-home testing is the lack of experimental control over the testing equipment and environment (Anwyl-Irvine et al., 2020a). However, there are some important advantages as well. For example, analyzing data collected from home samples facilitates the examination of

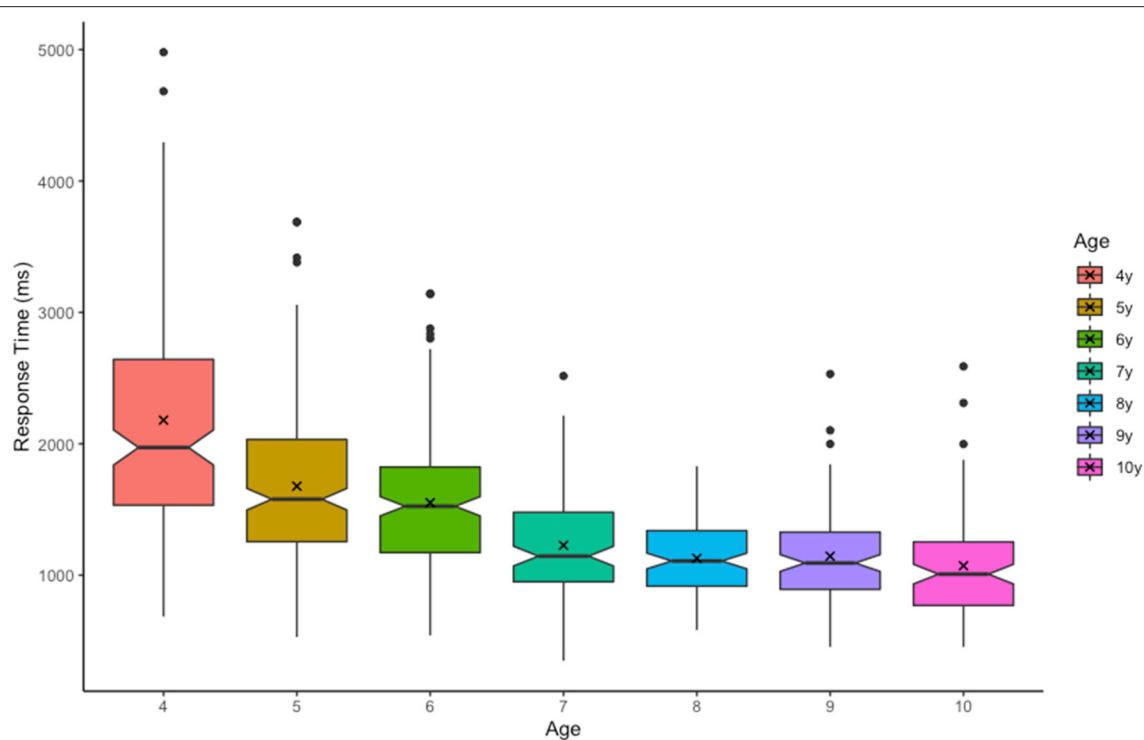


FIGURE 2 | Trial response times (ms) by age. Boxplot edges represent upper and lower quartiles, notches represent the 95% confidence interval of the median (center line), and 'X' represents the mean.

TABLE 2 | Pearson Bivariate correlation table of task and test environment factors. Significant effects indicated with (*).

| | Age | Trial count | Response time | Resolution | Response mode | Max K |
|---------------|----------|-------------|---------------|------------|---------------|----------|
| Age | 1 | 0.230** | −0.565** | −0.010 | −0.105 | 0.579** |
| Trial Count | 0.230** | 1 | −0.311** | 0.000 | 0.020 | 0.059 |
| Response Time | −0.565** | −0.311** | 1 | 0.120 | 0.199** | −0.419** |
| Resolution | −0.010 | 0.000 | 0.120 | 1 | 0.409** | 0.101 |
| Response Mode | −0.105 | 0.020 | 0.199** | 0.409** | 1 | −0.029 |
| Max K | 0.579** | 0.059 | −0.419** | 0.101 | −0.029 | 1 |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

often ignored task specifics such as the size of the screen (width in pixels), method of response (1 = touchscreen, 2 = keyboard, 3 = mouse) and their influence on VWM capacity estimates. Results of a correlation analysis revealed that neither screen size ($r = 0.101$) nor response mode ($r = -0.029$) were related to VWM capacity, though screen size and response mode were highly correlated, $r = 0.409$, $p < 0.001$ (Table 2). Response mode was also positively correlated with response time ($r = 0.199$, $p = 0.001$), revealing that children responded most quickly when using touchscreen devices (both computers and tablets). Several other significant relations were observed, most notably between response time and max K ($r = -0.419$, $p < 0.001$), with faster responding associated with higher capacity estimates, though age may have been an important driver of this effect.

Assessing Capacity Across Multiple Set Sizes

Although Pashler's capacity estimate is convenient and easily interpreted, using this equation with child populations poses some unique challenges. One such challenge occurs when hit rates are lower than false alarm rates. In these cases, Pashler's equation will produce a negative value that is uninterpretable. For example, one 5-year-old child in our sample had the following capacity estimates for array sizes 1 through 4, respectively: 1, 1.78, −0.86, and .44. There are two things to notice. First, this child had a negative value for array size 3 (−0.86), however estimates for array sizes 1 and 2 appear valid. Given these negative values were rare ($n = 9$ of 1,051 cells) we treated them as missing data and removed them from the analysis. The second thing to notice, is

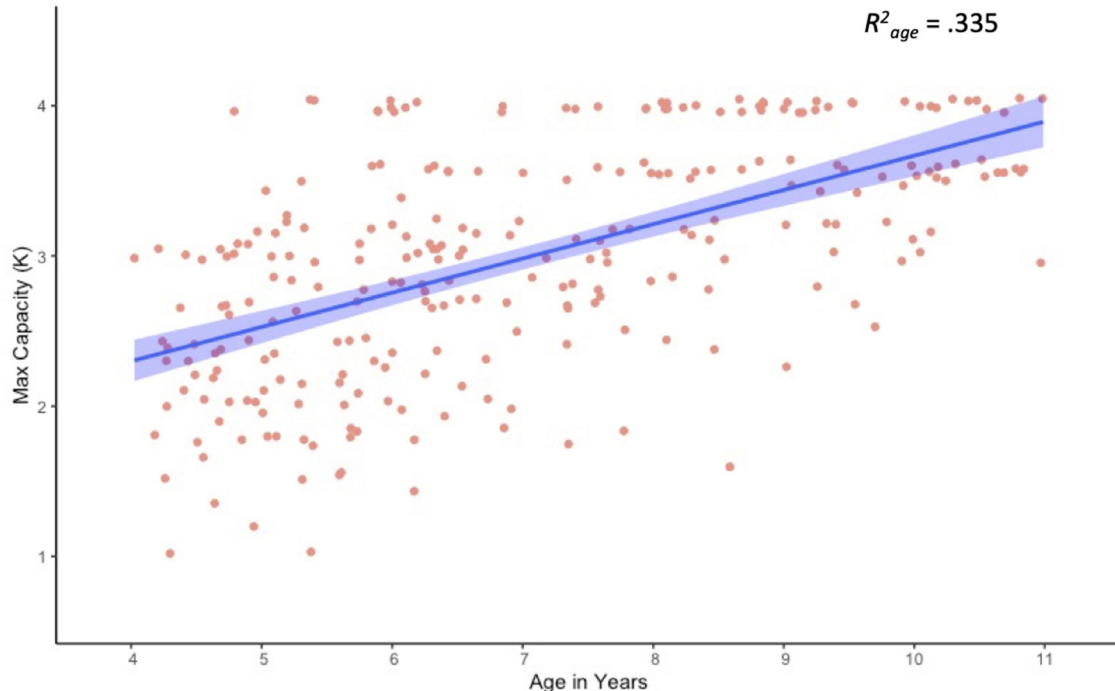


FIGURE 3 | Scatter plot and linear trend for visual working memory capacity (max K) as a function of age.

that the capacity estimate for array size 4 is *smaller* than estimates for array size 2 and even array size 1. We believe this may occur when children become overwhelmed by the memory demands for a given array and disengage from the task. There is some neurophysiological evidence to support this (Fukuda et al., 2010; Reyes et al., 2020; McKay et al., 2021). If this is the case, then the array size that produces maximum capacity (i.e., the *optimal array size*) should vary by age, with younger children reaching maximum capacity for smaller array sizes, and older children reaching maximum capacity for large array sizes independent of capacity estimates. An examination of the raw data clearly reveals such a trend (Figure 4), with younger children showing apparent capacity regressions at higher array sizes (Figure 5).

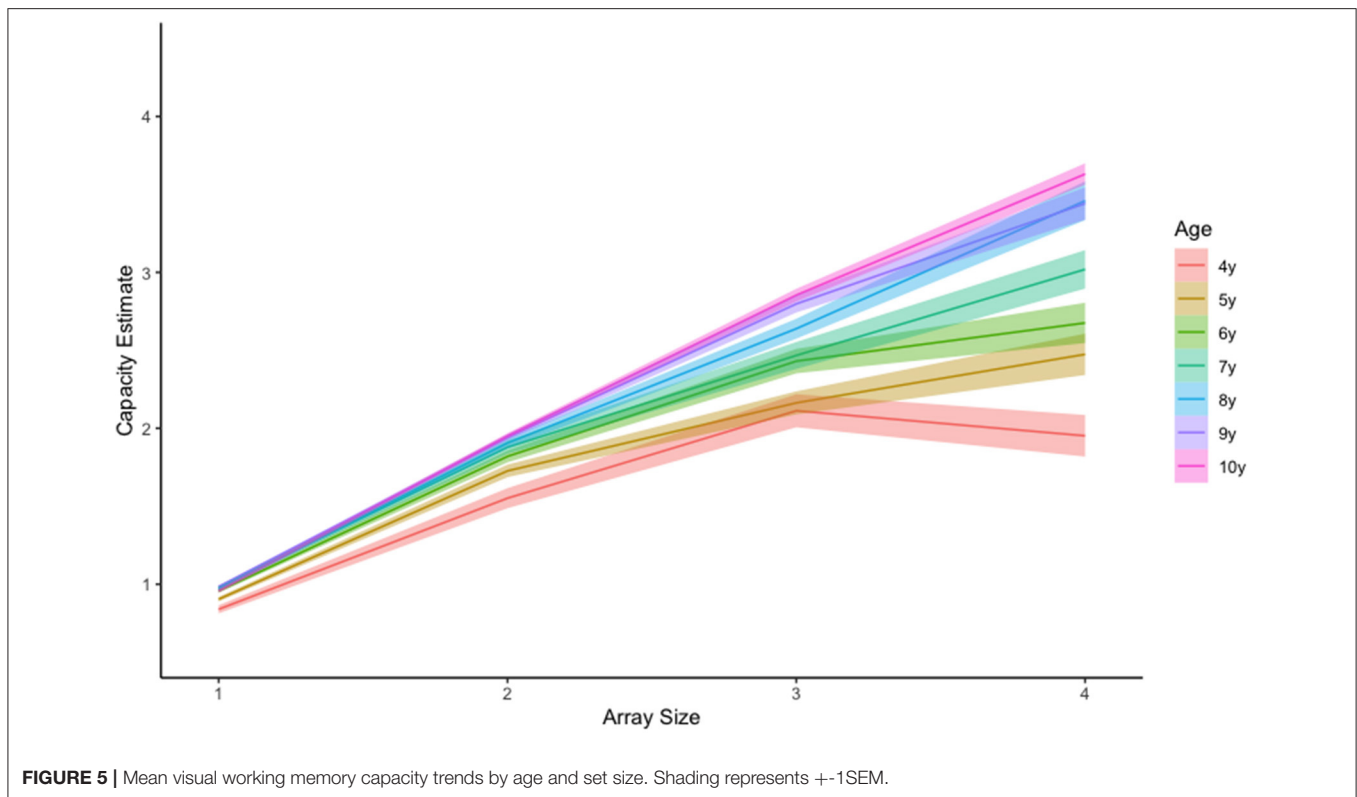
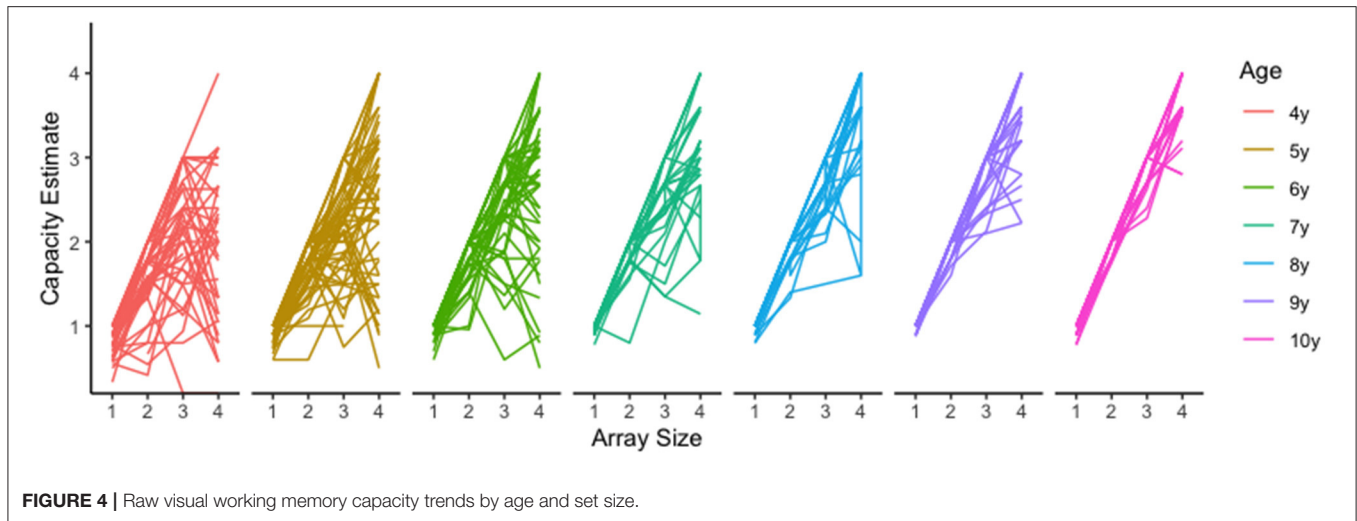
Do Large Arrays Disproportionately Hinder VWM Performance for Younger Children?

To determine if large array sizes resulted in underestimation of capacity for our young participants, we conducted a linear mixed effect (LME) analysis using R (R Core Team, 2020) with package lme4 (Bates et al., 2015). LME analyses are robust to missing data, and can handle the interdependence of capacity estimates across array size (Singmann and Kellen, 2019). This approach allowed us to calculate the extent to which capacity estimates increased with increasing array size for each age. We included fixed effects of array size and age and random participant-level effects in our baseline model (i.e., random intercept). Based on the observation that capacity varied with age (Figure 4),

we additionally included an array size by age interaction. This addition significantly improved model fits, $\chi^2(18, N = 1,051) = 199.93, p < 0.001$.

Effect estimates from our full LME model are presented in Table 3, and estimated marginal means are presented in Table 4. Age and array size were dummy coded so that the intercept reflects mean capacity for our reference group (4-year-olds at array size 1), and estimates reflect deviations from reference. Results for age were not significant, suggesting that despite small differences in array size 1 estimates (e.g., $K = 0.83$ at 4 years versus $K = 0.96$ at 10 years) all ages performed at ceiling for array size 1. However, results for array sizes 2–4 varied markedly by age. For example, though all ages had significant array size 4 effects, only 8- to 10-year-olds demonstrated significant array size 3 effects, with only 9- and 10-year-olds showing additional marginal effects for array size 2. This makes sense, as the slope of the regression line for array size should increase as overall capacity estimates increase (Figure 3).

To assess these patterns more directly, we conducted follow-up contrast analyses for each age (R package: emmeans v1.5.5-1) using estimated marginal means derived from our LME model (Searle et al., 1980). Significant non-linear trends would suggest that capacity estimates peaked for smaller array sizes, then regressed for larger array sizes. Results revealed significant quadratic trends for our four youngest ages: 4 years, $t_{(820)} = -6.551, p < 0.001$, 5 years, $t_{(817)} = -4.389, p < 0.001$, 6 years, $t_{(812)} = -5.007, p < 0.001$ and 7 years, $t_{(812)} = -2.242, p = 0.025$. These findings highlight 4–7 years as an ideal age at which



to identify and track individual differences, and underline the importance of including smaller array sizes to catch maximum capacity performance for younger children. Although we see a great deal of variability in our youngest participants, performance for 8-, 9-, and 10-year-olds did not appear to differ. This observation coupled with relatively large capacity estimates for these older children, suggests that VWM capacity improvements may have slowed by 8-years-of age, approaching adult capacity of around 3–4 items (Rouder et al., 2011; Zhang and Luck, 2011).

DISCUSSION

Children ages four through 10 were tested in an unsupervised, online change-detection task. Results from this paper highlight several novel benefits of online testing. For example, online approaches are quick, have compliance rates comparable to lab-based techniques, and appear to provide accurate results on par with lab-based approaches. In addition, online testing may increase diversity of the sample, facilitate testing across a wide array of ages, and allow for testing

TABLE 3 | Estimates and model fits for predictors of visual working memory capacity. Significant effects indicated with (*).

| Model | Source | Estimates | SE | df | t | p |
|----------------|-------------------|------------|---------------|--------------|-----------|-----------------|
| Full model | Intercept | 0.834 | 0.075 | 926.188 | 11.168 | <0.001*** |
| | Array size 2 | 0.717 | 0.092 | 791.429 | 7.754 | <0.001 |
| | Array size 3 | 1.276 | 0.094 | 797.340 | 13.611 | <0.001*** |
| | Array size 4 | 1.115 | 0.094 | 797.340 | 11.898 | <0.001*** |
| | 5y | 0.068 | 0.099 | 939.852 | 0.688 | 0.492 |
| | 6y | 0.123 | 0.102 | 914.485 | 1.206 | 0.228 |
| | 7y | 0.149 | 0.115 | 902.258 | 1.294 | 0.196 |
| | 8y | 0.141 | 0.116 | 892.237 | 1.222 | 0.222 |
| | 9y | 0.150 | 0.117 | 911.148 | 1.282 | 0.200 |
| | 10y | 0.124 | 0.121 | 910.477 | 1.026 | 0.305 |
| | Array size 2 *5y | 0.106 | 0.123 | 790.210 | 0.866 | 0.387 |
| | Array size 3 *5y | -0.022 | 0.124 | 794.534 | -0.176 | 0.861 |
| | Array size 4 *5y | 0.450 | 0.124 | 796.411 | 3.619 | <0.001*** |
| | Array size 2 *6y | 0.146 | 0.126 | 790.117 | 1.154 | 0.249 |
| | Array size 3 *6y | 0.198 | 0.127 | 793.346 | 1.556 | 0.120 |
| | Array size 4 *6y | 0.604 | 0.127 | 793.346 | 4.746 | <0.001*** |
| | Array size 2 *7y | 0.183 | 0.142 | 789.795 | 1.285 | 0.199 |
| | Array size 3 *7y | 0.213 | 0.143 | 792.350 | 1.486 | 0.138 |
| | Array size 4 *7y | 0.925 | 0.143 | 792.350 | 6.466 | <0.001*** |
| | Array size 2 *8y | 0.219 | 0.142 | 789.795 | 1.539 | 0.124 |
| | Array size 3 *8y | 0.393 | 0.143 | 792.350 | 2.747 | 0.006** |
| | Array size 4 *8y | 1.374 | 0.143 | 792.350 | 9.604 | <0.001*** |
| | Array size 2 *9y | 0.241 | 0.145 | 789.749 | 1.659 | 0.098 |
| | Array size 3 *9y | 0.538 | 0.146 | 792.208 | 3.688 | <0.001*** |
| | Array size 4 *9y | 1.343 | 0.146 | 792.208 | 9.204 | <0.001*** |
| | Array size 2 *10y | 0.276 | 0.150 | 789.676 | 1.838 | 0.066 |
| | Array size 3 *10y | 0.620 | 0.151 | 791.979 | 4.112 | <0.001*** |
| | Array size 4 *10y | 1.557 | 0.151 | 791.979 | 10.329 | <0.001*** |
| AIC | | BIC | LogLik | Chisq | df | p |
| Baseline model | | 1618.60 | 1678.10 | -797.30 | - | - |
| Full model | | 1454.70 | 1603.40 | -697.34 | 199.930 | 18.00 <0.001*** |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

across regions, or even countries. Other benefits of this approach include reduced resource and infrastructure demands, increased testing speed (~300 participants tested around 6 months vs. 2–3 years for in-lab testing), and the ability to allow maximum flexibility for parents and children, so that sessions may be timed when participants are maximally attentive.

Although there are several challenges to testing online, we did not find them to be unsurmountable. For example, ensuring that participants (not the parents) completed the assessments could be handled by capturing periodic facial images during testing, something that is possible with most browser-based experimental software suites. This may be particularly important if the task is being advertised broadly and compensation is provided. Although we did not collect participant video in our sample, we limited participation to families in our local area with whom we had a prior relationship, either as participants

in our own lab or in our departmental colleagues' labs. In addition, pre-screening the data prior to analysis can help identify suspicious data (e.g., response times too quick or performance too high). All tasks should be piloted in-lab to help develop expectations for performance, and to identify any issues with the task, or with child and/or parent understanding of the task.

Our results revealed several insights regarding at-home testing, such as the importance of tracking as many environment variables as possible. Although we found no evidence that screen size and response mode impacted VWM capacity estimates, it is possible that exceptionally large or small screens might still be problematic. We did find evidence that response mode influenced the speed of responding, which might be an issue for speeded designs or designs that require some sort of response inhibition (e.g., flanker or go/no go tasks). Some of our findings were not unique to online testing, such as the finding of slower response times for younger kids, and larger arrays sizes (older children only).

In addition to demonstrating the validity of unsupervised online testing approaches, our results also produced several novel insights regarding the development of VWM from 4 through 10 years of age. First, our results produced capacity estimates that are comparable to published lab-based estimates (Cowan et al., 2005; Riggs et al., 2006; Simmering, 2016), suggesting this approach to be a viable alternative requiring a fraction of the resources necessary for lab-based tasks. In addition, we found capacity increased significantly with age, reaching near-adult levels by around 8-years-of age (Figure 3). Our analysis also revealed evidence of substantial performance variability from 4- to 7-years-of-age (Figure 4), potentially highlighting assessment points for longer-term individual difference studies, as well as possible targets for memory intervention. Given the ease of online testing and the importance of VWM to several aspects of math and cognitive performance (Jarvis and Gathercole, 2003; Bull, 2008; Tsubomi and Watanabe, 2017; Giofrè et al., 2018; Allen et al., 2019; Chan and Wong, 2019; Kytälä et al., 2019; Carr et al., 2020), adding a quick at-home assessment as part of a school, medical, or lab assessment might provide a more detailed developmental profile.

On Estimating Capacity in Children

One of our most important findings was the demonstration of an interaction between array size and capacity estimation, especially for our youngest participants. Whereas, our older participants appeared able to perform consistently regardless of array size, our youngest participants seemed to disengage for larger arrays, resulting in estimates that were often lower than estimates obtained from smaller arrays. This is evidenced visually in our raw data (Figure 4), and statistically in our finding of significant quadratic trends for our 4- through 7-year-olds. These errors may have been purposeful (i.e., sample array perceived as too difficult resulting in a random guessing strategy), or they may have occurred after earnest attempts to respond accurately. If an explicit guessing strategy

was employed for larger array sizes, we would expect mean response times to be negatively correlated with array size. A correlation analysis on the raw data revealed this may be the case, with 4- through 7-year-olds demonstrating a small but significant *negative* correlation between response time and array size, $r = -0.061$, $p = 0.031$, and 8- through 10-year-olds revealing a small but significant *positive* correlation, $r = 0.093$, $p = 0.032$.

The finding of slightly faster response times for large array sizes suggests that at least some of our youngest participants may have resorted to guessing strategies when the demands of the array exceeded memory capacity, attentional resources, or some combination of the two. This is consistent with previous work demonstrating that children have sufficient metacognitive awareness to know when they have successfully encoded a to-be-remembered event, and when they have not (Applin and Kibbe, 2020). However, it is also possible that this drop in performance for set size 4 arrays may be the result of *catastrophic forgetting*, or the inability to encode any array items when capacity is exceeded. For example, in manual search tasks, 12- and 14-month-old infants appear unable to detect the difference between hiding events involving two vs. four balls, despite successfully detecting the difference between two vs. three balls (Feigenson and Carey, 2003). Importantly, this effect may have been partially driven by perceptual similarity, as it is largely ameliorated when four differently colored balls are used (Zosh and Feigenson, 2012). Given the older participant ages tested here and our use of highly discernably circle colors, it seems unlikely that the drop in performance for large arrays is the result of catastrophic forgetting.

Although adult researchers have proposed avoiding small array sizes to reduce the likelihood of underestimation (Morey, 2011), our results suggest that using large array sizes might also underestimate capacity, particularly for our youngest participants. Without a doubt, probabilistic and Bayesian approaches to capacity estimation are more sophisticated and can better account for high false alarm rates present in our young samples. However, these analysis techniques are not as readily adapted to online calculation or quick assessment for individual participants. We believe using a variety of array sizes works well as long as assessments are based on either *maximum* capacity across array sizes, or a holistic assessment of capacity as a *function of array size*. It is possible that reducing the number of large array sizes would increase number of trials young children complete, but those benefits would have to be weighed against the possible cost of underestimating capacity due to ceiling effects for higher performing children. If the goal of the assessment is to identify general working memory ability, a more desirable metric might be the *array size* at which a child reaches maximum capacity, or the *optimal array size*. This metric incorporates both a *quantitative* capacity estimate (i.e., maximum capacity) and a *qualitative* attentional estimate (i.e., maximum array size a child can tolerate before disengagement).

In conclusion, results presented here demonstrate the feasibility of effective and accurate at-home assessments of VWM, and provide novel insights into the influence of factors

TABLE 4 | Estimated marginal means based on best-fitting LME model (full model).

| Age | Set size | Mean | SE | df | Lower CI | Upper CI |
|----------|----------|-------|-------|-----|----------|----------|
| 4 years | 1 | 0.834 | 0.076 | 953 | 0.686 | 0.983 |
| | 2 | 1.551 | 0.075 | 944 | 1.404 | 1.698 |
| | 3 | 2.110 | 0.077 | 961 | 1.960 | 2.260 |
| | 4 | 1.949 | 0.077 | 961 | 1.799 | 2.100 |
| 5 years | 1 | 0.902 | 0.066 | 940 | 0.773 | 1.031 |
| | 2 | 1.725 | 0.066 | 940 | 1.596 | 1.854 |
| | 3 | 2.156 | 0.066 | 947 | 2.026 | 2.286 |
| | 4 | 2.467 | 0.067 | 960 | 2.335 | 2.599 |
| 6 years | 1 | 0.957 | 0.070 | 927 | 0.819 | 1.095 |
| | 2 | 1.820 | 0.070 | 927 | 1.682 | 1.958 |
| | 3 | 2.431 | 0.070 | 927 | 2.293 | 2.569 |
| | 4 | 2.676 | 0.070 | 927 | 2.538 | 2.814 |
| 7 years | 1 | 0.983 | 0.089 | 910 | 0.809 | 1.158 |
| | 2 | 1.883 | 0.089 | 910 | 1.708 | 2.057 |
| | 3 | 2.472 | 0.089 | 910 | 2.297 | 2.646 |
| | 4 | 3.024 | 0.089 | 910 | 2.849 | 3.198 |
| 8 years | 1 | 0.976 | 0.089 | 892 | 0.800 | 1.151 |
| | 2 | 1.911 | 0.089 | 892 | 1.736 | 2.087 |
| | 3 | 2.644 | 0.089 | 892 | 2.469 | 2.820 |
| | 4 | 3.465 | 0.089 | 892 | 3.289 | 3.640 |
| 9 years | 1 | 0.985 | 0.091 | 927 | 0.805 | 1.164 |
| | 2 | 1.942 | 0.091 | 927 | 1.762 | 2.121 |
| | 3 | 2.798 | 0.091 | 927 | 2.619 | 2.978 |
| | 4 | 3.442 | 0.091 | 927 | 3.263 | 3.622 |
| 10 years | 1 | 0.959 | 0.097 | 927 | 0.769 | 1.148 |
| | 2 | 1.951 | 0.097 | 927 | 1.761 | 2.141 |
| | 3 | 2.854 | 0.097 | 927 | 2.665 | 3.044 |
| | 4 | 3.631 | 0.097 | 927 | 3.441 | 3.820 |

such as array size, screen size, and response mode. Results additionally highlight numerous benefits for unsupervised at-home testing, from substantially increasing sample diversity (e.g., SES, race, ethnicity) to enabling large-scale geographically unconstrained population surveys at a relatively low cost. We have also found that allowing participants the flexibility to pick optimal test times increases compliance, decreases stress, and contributes to improved data quality and representativeness. Although this approach may not be useful for tasks that require closely monitored speeded approaches, it seems quite appropriate for change-detection tasks. Future work will be conducted to test older ages and broaden our participant pool geographically to include underrepresented regions and populations. It is our hope that approaches like the one presented here may help identify regional, cultural, and socioeconomic influences that affect VWM development and general cognitive outcomes.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article are available for download at https://osf.io/2b8zg/?view_only=44f50ae4514d415c8da887c53431fd14.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Tennessee IRB #17-03545. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

REFERENCES

- Allen, K., Higgins, S., and Adams, J. (2019). The relationship between visuospatial working memory and mathematical performance in school-aged children: a systematic review. *Educ. Psychol. Rev.* 31, 509–531. doi: 10.1007/s10648-019-09470-8
- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., and Evershed, J. K. (2020a). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behav. Res. Methods*. doi: 10.3758/s13428-020-01501-5. [Epub ahead of print].
- Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020b). Gorilla in our midst: an online behavioral experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x
- Applin, J. B., and Kibbe, M. M. (2020). Young children monitor the fidelity of visual working memory. *J. Exp. Psychol.* doi: 10.1037/xlm0000971. [Epub ahead of print].
- Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67:1–51. doi: 10.18637/jss.v067.i01
- Bays, P. M., and Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science* 321, 851–854. doi: 10.1126/science.1158023
- Bull, R. (2008). Short-term memory, working memory, and executive functioning in preschoolers: longitudinal predictors of mathematical achievement at age 7 years. *Dev. Neuropsychol.* 33, 205–228. doi: 10.1080/87565640801982312
- Buss, A. T., Ross-Sheehy, S., and Reynolds, G. D. (2018). Visual working memory in early development: a developmental cognitive neuroscience perspective. *J. Neurophysiol.* 120, 1472–1483. doi: 10.1152/jn.00087.2018
- Carr, M., Horan, E., Alexeev, N., Barsed, N., Wang, L., and Otumfuor, B. (2020). A longitudinal study of spatial skills and number sense development in elementary school children. *J. Educ. Psychol.* 112, 53–69. doi: 10.1037/edu0000363
- Chan, W. W. L., and Wong, T. T. Y. (2019). Visuospatial pathways to mathematical achievement. *Learn. Instruct.* 62, 11–19. doi: 10.1016/j.learninstruct.2019.03.001
- Cowan, N. (2001). The magical number 4 in short term memory. A reconsideration of storage capacity. *Behav. Brain Sci.* 24, 87–186. doi: 10.1017/S0140525X01003922
- Cowan, N., Elliott, E. M., Sauls, S. J., Morey, C. C., Mattox, S., Hismjatullina, A., et al. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cognit. Psychol.* 51, 42–100. doi: 10.1016/j.cogpsych.2004.12.001
- Feigenson, L., and Carey, S. (2003). Tracking individuals via object-files: evidence from infants' manual search. *Dev. Sci.* 6, 568–584. doi: 10.1111/1467-7687.00313
- Fukuda, K., Vogel, E., Mayr, U., and Awh, E. (2010). Quantity, not quality: the relationship between fluid intelligence and working memory capacity. *Psychon. Bull. Rev.* 17, 673–679. doi: 10.3758/17.5.673
- Gathercole, S. E., and Baddeley, A. D. (1993). "Working memory and language," in *Essays in Cognitive Psychology* (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc). <http://proxy.lib.utk.edu:90/login?url=http://search.proquest.com/docview/618406033?accountid=14766>
- Gathercole, S. E., Pickering, S. J., Ambridge, B., and Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Dev. Psychol.* 40, 177–190. doi: 10.1037/0012-1649.40.2.177
- Giofrè, D., Donolato, E., and Mammarella, I. C. (2018). The differential role of verbal and visuospatial working memory in mathematics and reading. *Trends Neurosci. Educ.* 12, 1–6. doi: 10.1016/j.tine.2018.07.001
- Jarvis, H., and Gathercole, S. E. (2003). Verbal and non-verbal working memory and achievements on national curriculum tests at 11 and 14 years of age. *Educ. Child Psychol.* 20, 123–140.
- Kyttälä, M., Kanerva, K., Munter, I., and Björn, P. M. (2019). Working memory resources in children: stability and relation to subsequent academic skills. *Educ. Psychol.* 39, 709–728. doi: 10.1080/01443410.2018.1562046
- Luck, S. J., and Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature* 390, 279–281. doi: 10.1038/36846
- McKay, C. A., Shing, Y. L., Rafetseder, E., and Wijeakumar, S. (2021). Home assessment of visual working memory in pre-schoolers reveals associations between behaviour, brain activation and parent reports of life stress. *Dev. Sci.* 1–14. doi: 10.1111/desc.13094
- Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *J. Math. Psychol.* 55, 8–24. doi: 10.1016/j.jmp.2010.08.008
- Oakes, L. M., Ross-Sheehy, S., and Luck, S. J. (2006). Rapid development of feature binding in visual short-term memory. *Psychol. Sci.* 17, 781–787. doi: 10.1111/j.1467-9280.2006.01782.x
- Pashler, H. (1988). Familiarity and visual change detection. *Percept. Psychophys.* 44, 369–378. doi: 10.3758/B.F.03210419
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
- Reyes, L. D., Wijeakumar, S., Magnotta, V. A., and Spencer, J. P. (2020). Localizing changes in the functional brain networks that underlie visual working memory in the first two years of life. *Neuroimage*. 219:116971. doi: 10.1016/j.neuroimage.2020.116971
- Riggs, K. J., McTaggart, J., Simpson, A., and Freeman, R. P. J. (2006). Changes in the capacity of visual working memory in 5- to 10-year-olds. *J. Exp. Child Psychol.* 95, 18–26. doi: 10.1016/j.jecp.2006.03.009
- Ross-Sheehy, S., and Eschman, B. (2019). Assessing visual STM in infants and adults: eye movements and pupil dynamics reflect memory maintenance. *Vis. Cogn.* 27, 78–92. doi: 10.1080/13506285.2019.1600089
- Ross-sheehy, S., Oakes, L. M., and Luck, S. J. (2003). The development of visual short-term memory capacity in infants. *Child Dev.* 74, 1807–1822. doi: 10.1046/j.1467-8624.2003.00639.x
- Rouder, J. N., Morey, R. D., Morey, C. C., and Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychon. Bull. Rev.* 18, 324–330. doi: 10.3758/s13423-011-0055-3
- Searle, S. R., Speed, F. M., and Milliken, G. A. (1980). Population marginal means in the linear model: an alternative to least squares means. *Am. Stat.* 34, 216–221. doi: 10.1080/00031305.1980.10483031
- Simmering, V. R. (2012). The development of visual working memory capacity during early childhood. *J. Exp. Child Psychol.* 111, 695–707. doi: 10.1016/j.jecp.2011.10.007
- Simmering, V. R. (2016). I. Working memory capacity in context: modeling dynamic processes of behavior, memory, and development. *Monogr. Soc. Res. Child Dev.* 81, 7–24. doi: 10.1111/mono.12249
- Simmering, V. R., and Perone, S. (2013). Working memory capacity as a dynamic process. *Front. Psychol.* 3, 1–26. doi: 10.3389/fpsyg.2012.00567
- Simmering, V. R., and Spencer, J. P. (2008). *Developing a magic number: the dynamic field theory reveals why visual working memory capacity estimates differ across tasks and development* (Doctor of Philosophy thesis). Iowa City, IA: University of Iowa. doi: 10.17077/etd.ugpho5mg
- Singmann, H., and Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. *New Methods in Cognitive Psychology*, 4–31. doi: 10.4324/9780429318405-2

AUTHOR CONTRIBUTIONS

SR-S, ER, and BE contributed to conception and design of the study. SR-S and ER performed the statistical analysis. SR-S wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

- Tsubomi, H., and Watanabe, K. (2017). Development of visual working memory and distractor resistance in relation to academic performance. *J. Exp. Child Psychol.* 154, 98–112. doi: 10.1016/j.jecp.2016.10.005
- Zhang, Q., Li, Y., Zhao, W., Chen, X., Li, X., Du, B., et al. (2020). ERP evidence for the effect of working memory span training on working memory maintenance: a randomized controlled trial. *Neurobiol. Learn. Memory* 167:107129. doi: 10.1016/j.nlm.2019.107129
- Zhang, W., and Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature* 453, 233–235. doi: 10.1038/nature06860
- Zhang, W., and Luck, S. J. (2011). The number and quality of representations in working memory. *Psychol. Sci.* 22, 1434–1441. doi: 10.1177/0956797611417006
- Zosh, J. M., and Feigenson, L. (2012). Memory load affects object individuation in 18-month-old infants. *J. Exp. Child Psychol.* 113, 322–336. doi: 10.1016/j.jecp.2012.07.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Ross-Sheehy, Reynolds and Eschman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Webcams, Songs, and Vocabulary Learning: A Comparison of In-Person and Remote Data Collection as a Way of Moving Forward With Child-Language Research

Giovanna Morini* and Mackensie Blair

Department of Communication Sciences and Disorders, University of Delaware, Newark, DE, United States

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Nicola A. Gillen,
University of Oxford, United Kingdom
Marianella Casasola,
Cornell University, United States

*Correspondence:

Giovanna Morini
gmorini@udel.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 29 April 2021

Accepted: 14 July 2021

Published: 09 August 2021

Citation:

Morini G and Blair M (2021)
Webcams, Songs, and Vocabulary
Learning: A Comparison of In-Person
and Remote Data Collection as a Way
of Moving Forward With
Child-Language Research.
Front. Psychol. 12:702819.
doi: 10.3389/fpsyg.2021.702819

This article evaluates a testing procedure for collecting eye-gaze data with toddlers and preschoolers during a word-learning task. We provide feasibility and precision data by comparing performance in an in-person version of the study (conducted under controlled conditions in the lab), with performance in a virtual version in which participants completed the testing procedure from home. Our data support the feasibility of collecting remote eye-gaze data with young children, and present it as a viable alternative for conducting developmental language research when in-person interactions with participants cannot take place. Additionally, we use this methodological approach to examine a topic that has gained popularity in recent years—the role of music and songs on vocabulary learning. We provide evidence suggesting that while songs may help increase attention during a particular task, greater attention does not lead to greater learning. In fact, preschoolers show improved word-learning performance for items that were trained in a spoken sentence compared to items that were trained in a song. This means that while songs may be beneficial for increasing child engagement, spoken sentences may be best for supporting deep level learning of language concepts.

Keywords: remote testing, word learning, eye-gaze measures, songs, toddlers, preschoolers

INTRODUCTION

Over the last 50 years we have seen important shifts toward new testing paradigms that would help shape theories of language acquisition. While initially, the study of child language had been restricted to the examination of early speech productions (Brown, 1973; Shatz, 1978), the introduction of new testing techniques, such as the Intermodal Preferential Looking paradigm (IPLP) (Golinkoff et al., 1987) would allow researchers to explore processes associated with language acquisition, even before children can produce words. The IPLP measures the speed and/or accuracy of children's looking patterns to objects on a screen, and since eye-gaze is an overt behavioral response that is present early in life, it does not rely heavily on motor control (Golinkoff et al., 2013). The IPLP has been used for decades in labs across the world, and has contributed to our understanding of critical skills within language acquisition such as word learning

(Hollich et al., 2000; Halberda, 2003; Newman et al., 2018) and word comprehension (Fernald et al., 2001; Swingley and Aslin, 2002; Houston-Price et al., 2007; Morini and Newman, 2019), in children as young as 6 months (Tincoff and Jusczyk, 1999; Bergelson and Swingley, 2012). But traditionally, this paradigm required participants to visit the lab, where children would be tested in a controlled environment (i.e., a quiet room with minimal distractions), using the same equipment across participants (i.e., the same screen, speakers, and video camera). Recently, unprecedented circumstances linked to the global pandemic have pushed researchers from across fields to explore new ways to collect data—as the majority of in-person testing has been halted. Many child-language researchers have turned to virtual methods as a way of accessing diverse participants, recruiting larger sample sizes, and continuing data collection in a way that remains pandemic-proof. However, many questions remain regarding the feasibility and sensitivity of data collected via remote testing. This is particularly true, when it comes to fine-grained measures such as eye-gaze and testing of young children who inherently have limited attention and cooperation spans.

As part of the present work, we developed a virtual version of the IPLP, and compared data collected in the lab under controlled conditions (pre-pandemic) to data collected virtually (during the pandemic) with children of the same age. This approach enabled us to examine a methodological aim, which focused on addressing some of the uncertainty surrounding the precision and feasibility of a remote approach. Part of the process of developing a virtual version of the IPLP involved deciding which type of language task to ask participants to complete. We chose to use a word learning task, in which participants were taught novel word-object pairings in two experimental conditions: in songs and in spoken sentences. This decision was motivated by the following factors: (i) in-person data collection for this task was underway in the lab, so we had available data that could be compared to that of children tested virtually, and (ii) little is known about the role that songs play on preschooler's ability to learn novel vocabulary items, which meant that we would have the opportunity to address a theoretical aim in addition to the methodological one.

In recent years music interventions and learning-through-song programs, including those that target vocabulary learning for children of various ages have increasingly gained popularity (Overland, 2017). Previous research examining the role of music on language learning has primarily focused on identifying shared learning mechanisms—for example, identifying similarities between music and language and the acquisition of skills across the two (Trehub and Trainor, 1993; Trehub, 2003; Brandt et al., 2012). However, there is limited work evaluating any direct benefits of music and song on the language acquisition process itself. This information can be particularly informative for caregivers, educators, and clinicians working with young children. Teaching words through songs is a practice that can be easily incorporated into everyday activities in a variety of settings (e.g., home, classroom) and that is, in fact, widely used. Though a popular practice, we have very little empirical data on the impact of music and song on language learning.

In fact, narrowing down a concrete definition of *what* music is and *how* its features might facilitate learning across domains has proven to be remarkably hard (Cross and Morley, 2008). Music has been described as a “universal feature of human cognition,” and it can be found universally across human cultures (Brandt et al., 2012). Music, like language, expresses rhythm, emotion, and meaning, and can help convey information in attention-grabbing ways, which might be especially useful for the learning process in young children (Simpson and Keen, 2009). There is considerable evidence suggesting that certain speech registers (e.g., infant-directed speech—IDS) are characterized by a slow speaking rate, high pitch, long vowels, greater rhythmicity and repetition (Stern et al., 1982, 1983; Fernald and Simon, 1984)—making this type of speech appear more “musical” compared to adult-directed speech (ADS) (Fernald, 1992). Furthermore, young children show a robust preference for IDS over ADS (Frank et al., 2020), and there is evidence suggesting that during the beginning stages of vocabulary learning, IDS may facilitate the acquisition and recognition of words (Thiessen et al., 2005; Singh et al., 2009; Ma et al., 2011). Similarly, certain forms of music (infant-directed versions of signing in particular) have overlapping features with IDS—including a slow tempo, high pitch, and repetition (Trainor et al., 1997; Trehub et al., 1997a,b; Trehub and Trainor, 1998). These shared characteristics would suggest that perhaps children's songs, like IDS, might facilitate vocabulary learning. Nevertheless, there has been an ongoing debate regarding differences and similarities in how young children process linguistic and musical features (Pinker, 1997; Jackendoff, 2009; Peretz, 2009).

One area that has been widely studied is the role of music and songs on attention. This is an important topic, given that attention is often described as a necessary early step in the learning process. Specifically, by relying on attention skills the learner is able to choose what information from the environment is relevant (and needs to be processed), and what information should be ignored because is not relevant to complete the task at hand (McDowd, 2007). Previous work with infants between the ages of 5 and 10 months suggests that hearing children's songs leads to greater engagement and sustains attention compared to hearing other types of auditory signals (e.g., other types of music, IDS, or ADS) (Trainor, 1996; Corbeil et al., 2016). In another study, infants (5.5–6.5 months) attended longer to videos of their mothers singing than videos of their mothers speaking, further supporting a preference for songs over speech (Nakata and Trehub, 2004). However, Corbeil et al. (2013) examined whether specific features included in songs (and speech) might guide infants' preference for the different types of auditory stimuli. They found that children did not show a particular preference for melodic features of music and song, and instead showed a preference for happier sounding stimuli. For example, infants preferred to listen to IDS over a hummed melody, as well as happy sounding infant-directed song over more neutral IDS. Furthermore, infants showed no preference between happy-sounding IDS and infant-directed song. The role of music and songs on attention has also been studied in slightly older children. Wolfe and Noguchi (2009) presented 5-year-old children with stories either in speech or in a song modality. Some children

heard the story with background auditory distractors, while others did not. Auditory distractors were presented in both modalities of story presentation. When distractors were present, participants were better able to recall information about the content of the story when the story was heard in a song, compared to the spoken condition. The authors concluded that music may increase selective attention and awareness in school-aged children. Taken together, these findings suggest that there is a robust attentional preference for songs over speech that has been documented in infancy and into early childhood. While the features that are driving this effect are not fully understood, there is some evidence suggesting that certain characteristics of the auditory signal (e.g., affect) might play a bigger role guiding infant's engagement than others (e.g., melodic changes alone).

It is important to note that showing preference for a particular auditory signal, does not necessarily translate to greater learning. When it comes to the role of music and song and its relation to learning in the language domain, existing findings are mixed, and they come primarily from studies with older children who were second-language (L2) learners (Salcedo, 2010; Ludke et al., 2014; Good et al., 2015; Busse et al., 2018). In one study Coyle and Gómez Gracia (2014) presented Spanish speaking 5-year-olds who were learning English as an L2 with lessons targeting specific English vocabulary words. These lessons were taught using a popular children's song "The wheels on the bus." The song was used to teach five target words. The children received three 30-min teaching sessions using this song. The sessions were structured as follows: the teacher first explained and identified the target words using a visual of the bus, then the teacher sang the song twice emphasizing the target words and their location (all words were part of the bus). Before each lesson children were asked to identify and produce the target vocabulary learned in the song. The authors found that children were better able to identify the target words receptively after each lesson in comparison to their performance before instruction. However, there was no change in their ability to produce the target words. These findings suggested that using a song to present novel target words facilitated receptive vocabulary, but did not lead to improved learning in expressive vocabulary. Another study with school-aged children between 10 and 14 years of age in Thailand examined incidental learning of vocabulary words in English (the participant's L2) by exposing participants to popular songs in English, and testing them on specific vocabulary words found in each of the songs (Pavia et al., 2019). The results indicated that the more the children were exposed to the songs, the better they were able to recall the target words within the songs. In addition to vocabulary learning, the use of music and songs has been found to enhance the acquisition of grammar skills in an L2. For example, Legg (2009) found that music aided 12–13-year-old students in French-learning classrooms during instruction of past tense verbs. Specifically, using a song to demonstrate and practice past-tense use led to higher scores at post-test than when a song was not used as part of the lessons.

Fewer studies have explored the role of music and song on language learning in young children's native language. Thiessen and Saffran (2009) presented infants (between 6.5 and 8 months) with a sequence of numbers either in spoken sentences or in

a song. After a familiarization period, infants were presented with the same sequence of numbers, or a novel sequence to test whether or not they had learned the original number pattern. Testing always occurred in speech, regardless of the modality of familiarization. Infants showed a preference for the novel string suggesting that they could differentiate it from the trained sequence, only when familiarization had occurred in song, but not when they had been trained in speech. Another study with 11-month-olds examined infants' ability to detect changes in phonetic and melodic information within songs (Lebedeva and Kuhl, 2010). When participants were familiarized with a consistent four note melody, they were able to detect a change in the sequence of notes. However, when they were familiarized with a four-syllable spoken non-sense word, they were not able to detect a change in syllable order. In a follow-up task the authors examined whether embedding the non-sense words in a consistent melody (i.e., a song) would improve infants' ability to detect the change in syllable order. They found that, in fact, there was an increase in phonetic recognition when the non-sense words were presented in the song context. Lastly, one electrophysiological study examined whether 10-month-old Dutch-learning infants could segment target words that were presented in a song or in a speech stream during familiarization, and whether one condition would lead to better recognition of those words when they were presented in continuous speech (Snijders et al., 2020). Analyses of event-related potentials (ERPs) suggested that there was no difference in segmentation abilities across the two conditions (i.e., infants segmented words during both speech and song familiarization). Furthermore, there was no evidence that children could recognize the familiarized words during test trials following either song or speech. In other words, there was no evidence of songs providing a facilitatory effect during this particular task.

Nevertheless, in the majority of the previous studies participants were not asked to learn word-object relations; instead, they were tested on their ability to recognize auditory patterns that were presented during familiarization/training (e.g., numbers, words). But in the real world, children must go beyond simply tracking auditory patterns to expand their vocabulary; they must learn relations between specific sound patterns and a concrete referent (Stager and Werker, 1997; Werker et al., 1998, 2002). To learn a word like "apple" from the utterance "look at the apple!" children must first segment the target word from the continuous stream of speech, they must then identify the referent that corresponds to the new word, next they must encode the sequence of phonemes that make up the word, and lastly store the new word-referent association so that it can be retrieved later on (Capone and McGregor, 2005; Gupta, 2005). Furthermore, these associations must be generated and stored relatively fast in order for vocabulary growth to occur at the speed that it does; that is, children's vocabulary increases rapidly and it is not the case that children spend months or even weeks learning a single word.

Taken together, previous work has supported the notion that music and songs can facilitate children's memory for verbal material, with evidence coming primarily from second language vocabulary acquisition. However, the findings are mixed and the "song advantage" appears to be specific to some tasks but not

others. Furthermore, there is limited data examining the role of songs on language development in young children's native language, and specifically the role of songs when it comes to acquiring novel word-object relations. Hence, additional research is needed to (i) confirm prior findings, and (ii) extend this work to vocabulary learning tasks that more closely resemble the word-learning process that young children face when acquiring words in the real-world.

The present study examined two main topics. As a first step, we aimed to investigate the efficacy and feasibility of a virtual version of the IPLP for studying word learning in young children. Additionally, we wanted to know whether training novel words through songs would lead to better acquisition of the word-object pairs compared to when words were trained using a spoken sentence. As part of the study, children were taught two new words that corresponded to novel objects. One of the words was trained using a spoken sentence produced in IDS prosody, while the second word was trained in a song. Children were then tested on their ability to recognize each item using a modified version of the Intermodal Preferential Looking Paradigm (Golinkoff et al., 1987). The overall design was identical to the one used by Schmale et al. (2012) and Newman et al. (2018) to examine word learning in children of a similar age. Participants completed the same task either *in-person*, or *virtually*, with the goal of answering the following questions:

- 1A) Can preschoolers successfully engage and provide codable usable data in a virtual IPLP task completed from home?
- 1B) Does the modality of the testing procedure (i.e., in-lab vs. remote testing) influence the pattern of results?
- 2A) Does the use of song result in different patterns of novel word learning compared to the use of spoken sentences?
- 2B) Does age mediate word learning accuracy in the spoken or song conditions?

METHODS

Participants

Our sample included a total of 59 typically-developing preschoolers, divided into two age groups: (i) 29–32 month-olds ($N = 38$), and (ii) 47–50 month-olds ($N = 21$). Within the 29–32 month-old group, 29 of them were White, 4 were African American, and 5 were of mixed race. Within the 47–50 month-old group, 18 of them were White, 1 was African American, 1 was Hispanic, and 1 was of mixed race. Additional descriptive information for both age groups is presented in **Table 1**. Based on parental report, participants were being raised in monolingual English-speaking homes, and had not been diagnosed with any disabilities. The younger age group was selected because it is one that has been previously tested using in-person versions of the IPLP during similar word-learning tasks (Schmale et al., 2011; Newman et al., 2018), and because it is an age-range in which children are rapidly expanding their lexical skills (Fenson et al., 1994). The second age group was included to see whether the virtual version of the IPLP could also be successfully used with slightly older children. The idea being that 47–50 month-olds have had more exposure to screens and electronic devices

TABLE 1 | Demographic information.

| Age group | | In-person | Virtual |
|-----------|----------------------------------|------------------------|------------------------|
| 29–32 | Sample size | $N = 19$ | $N = 19$ |
| | Gender | Male = 4 | Male = 9 |
| | Age | $M = 30.47, SD = 1.14$ | $M = 30.36, SD = 1.04$ |
| | Caregiver's education (in years) | $M = 18.11, SD = 2.56$ | $M = 16.67, SD = 2.14$ |
| 47–50 | Sample size | $N = 6$ | $N = 15$ |
| | Gender | Male = 3 | Male = 6 |
| | Age | $M = 48.68, SD = 0.93$ | $M = 48.72, SD = 1.07$ |
| | Caregiver's education (in years) | $M = 15.83, SD = 2.71$ | $M = 17.8, SD = 2.18$ |

(Certain and Kahn, 2002), and hence they might find sitting in front of a computer at home less novel/engaging, which might affect remote task performance. Additionally, 4-year-olds might approach the word-learning task differently. For example, they are now singing songs themselves regularly, and might rely more heavily on features of the song (e.g., the melody) during encoding of the word-object relations, which would lead to different patterns of performance compared to the toddlers in the younger group.

Half of the participants ($N = 19$) in the 29–32 month-old group completed the study *in-person* using an in-lab version of the IPLP, prior to in-person data collection being suspended due to the COVID-19 pandemic. The other half was tested with a *virtual* version of the same task, and participants were recruited until we could match the sample size of the in-person group. Most of the participants in the 47–50 month-old group completed the study in the *virtual* modality ($n = 15$), and only a small number was able to complete testing *in person* ($n = 6$). Our initial goal was to test a total of 19 participants in the older group (to match the sample size that was used for the younger groups). However, two additional participants were scheduled by lab staff for the older group during the recruitment process, and since the appointments were completed, we decided to include them in the final sample. As part of the inclusionary criteria for children completing the task in-person, families needed to be able to visit the lab to complete a 30-min testing session. To be included in the virtual testing, participants needed to have access to a computer with a webcam and a screen size of 12 inches or greater, as well as a reliable internet connection.

Stimuli

Two pairs of novel objects (4 objects total) were used to create the visual stimuli. In the videos the objects were waved back and forth to maintain participants' attention. Pairs of objects were matched for material (i.e., all were made of wood), size, and anticipated salience. Each object was a different solid color.

A female native speaker of American English recorded the auditory stimuli. The stimuli consisted of training sentences and test sentences. Training sentences were either *spoken* using IDS

prosody or produced in a *song* to the melody of “Old Mac Donald Had A Farm” (see **Figure 1**). The sentences included the carrier phrase (“Look! It’s a _____. Wow, it’s a _____. Do you see it? A _____”) followed by a target word. A total of four novel target words (to match each of the four novel objects) were presented during the study. All novel-words were one syllable long, and followed English phonotactic rules (e.g., *doop*, *neff*, *shoon*, *fim*). To ensure that the intelligibility of the context phrases was comparable across trials of the same condition, one token of each carrier phrase was selected and used for each target word. Additionally, three tokens of each target word per condition were selected (one for each of the 3-sentence carrier phrase), and cross-spliced into the sentences in the carrier phrase sequence.

Test sentences were produced by the same female speaker, and instructed children to look at one of the two objects on the screen (“Look at the ____! Do you see the ____? Where is that ____? ____!”). Note that this sequence ended with the final word presented in isolation, which was not the case for the training phrases. Additionally, all test phrases were produced in spoken sentences using IDS prosody. Once again, recordings of the different target words were cross-spliced into the same recording of the carrier phrase.

The onset of the first repetition of the target word occurred 1.4 s after the onset of the phrase; this was true for both training and testing trials. All trials were matched for amplitude and were 7.5 s in duration. Recordings were created using a Shure MV51 microphone at a 44.1 kHz sampling rate, 16-bits precision, inside a sound-attenuated booth. A sample video of the experimental task is available in a public scientific repository for this project (<https://osf.io/pfazg/>).

Procedure

In-person

Participants in the in-person group sat on their caregiver’s lap inside a sound-attenuated booth. A 43” LCD TV screen was positioned ~5.5 feet from the participant and was used to display the videos of the novel objects on a white background. The auditory stimuli were presented through a center speaker located above the TV. Caregivers were asked either to wear headphones and listened to masking music or close their eyes during the task, to avoid biasing children’s responses. An experimenter was able to see the caregiver and child with a camera throughout the duration of the study to ensure that the caregiver’s headphones remained on or their eyes stayed close. The testing paradigm was divided into four testing blocks: two in the song training condition and two in the spoken training condition (see **Figure 2** for an example of the presentation of stimuli in a block). Each block began with a baseline trial in which an object pair was presented on the screen without accompanying auditory stimulus. Baseline trials were included to allow us to check for object biases. After these silent trials, three training trials were then presented. During these trials a single object appeared in the center of the screen and was accompanied by sentences presented either in the song or the spoken condition. Testing for each of the word pairs occurred immediately after the training trials within each block. Blocks 1 and 2 each taught a new word: one in the song, and one in the spoken condition, and then tested

that learning on the two test trials, with one trial asking for the trained object and the other asking participants to look at a novel object. Blocks 3 and 4 were an exact repetition of the first two blocks. The idea behind this design is that if children have learned the trained word-object relation, they should look longer at the trained object when it is requested. Additionally, based on the principle of mutual exclusivity (Merriman and Bowman, 1989), which assumes that objects have a single label, children should look longer at the untrained object when they are asked to look at the item that was not trained. This means that the two test trials within each block assessed successful learning of the trained word-object pairing via mutual exclusivity for the untrained test, and through direct recall of the information provided in the training trials for the trained test. This type of approach is necessary to control for trained object preferences that may arise as a result of seeing the trained object more times during the training phase. In order to be included in the final sample, participants needed to have completed (i.e., had usable data for) at least one block in each of the experimental conditions.

The following parameters were counterbalanced across participants: (i) which word was presented as the trained word, (ii) which type of test trial, trained or novel, was presented first at test, (iii) whether the song or the spoken condition appeared during the first and third blocks or the second and fourth blocks, and (iv) which object received which label. Additionally, the left vs. right position of objects on the screen was counterbalanced across blocks for each participant. An 8-s video of a dancing Elmo cartoon on a black background was included between trials to maintain children’s attention. Since the trial videos had a white background and the attention-getter video had a black background, this led to changes in brightness detected by the camera that could be used to accurately identify the beginnings and ends of trials in the videos of the participants that were generated during testing. All trials had the same set duration (7.5 s) and automatically started after the Elmo attention-getter video was done playing. Visual stimuli appeared 0.4 s prior to the auditory stimulus, and the trials played uninterrupted from beginning to end. The Behavioral Infant and Toddler Testing System (BITTSy) (Newman et al., 2021) was used to control the stimulus presentation, and a video camera inside the testing booth was used to record videos of participants completing the task for later coding.

Virtual

Participants in the virtual group completed the study from home via a Zoom video call. Caregivers were asked to find a quiet room in the home and to try to avoid having any distractors present during the appointment (e.g., turning off the TV or music in the background). A detailed written testing protocol, which included step-by-step instructions to guide the appointment, as well as verbal scripts to explain the procedure to the families was generated and used for every testing session. This document is available in a public scientific repository (<https://osf.io/pfazg/>). This ensured that there was consistency across appointments, and made it possible to test families with varying levels of technical expertise. Experimenters received training on how to use Zoom and how to trouble-shoot issues that may arise



FIGURE 1 | Sample of auditory stimuli heard during training trials in the song condition.



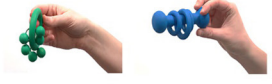

| Trial # | Trial Type | Audio | Visual | Prediction |
|---------|---|---|--|--|
| 1 | Baseline | Silence |  | Participants are expected to look at each object equally, approximately 50% of the time each |
| 2, 3, 4 | Training (song or spoken depending on block) | Look! It's a <i>doop</i> . Wow, it's a <i>doop</i> . Do you see it? A <i>doop</i> ! |  | NOTE: This trial repeats 3 times to allow children to learn the word-object pairing |
| 5 | Trained Test (spoken) | Look at the <i>doop</i> ! Do you see the <i>doop</i> ! Where is that <i>doop</i> ? <i>doop</i> ! |  | If learning occurs, children should look at the trained object (right) longer than the untrained object. |
| 6 | Untrained Test (spoken) | Look at the <i>fim</i> ! Do you see the <i>fim</i> ! Where is that <i>fim</i> ? <i>fim</i> ! |  | Children are expected to look longer to the untrained object (left), based on mutual exclusivity. |

FIGURE 2 | An example of the presentation of stimuli in a block.

during the appointments across different operating systems (e.g., Windows and Mac). A back-up experimenter (listed as “co-host” in the Zoom call) was always present, in case there were internet connectivity issues with the lead tester (i.e., this would avoid the call being dropped if one of the experimenters got disconnected).

The virtual appointment started with a light, camera, and audio check. Using the chat function in Zoom, the experimenter provided caregivers with a link to a 30-s video of a spinning wale with music playing (this video is available in the public scientific repository for this project: <https://osf.io/pfazg/>). The background color of the video changed from black to white every 5 s, allowing the experimenter to see if the changes in brightness (e.g., from black to white) were detectable via the webcam (as this would be used to identify beginnings and ends of trials during coding). If the contrasts were not noticeable, the experimenter

asked the caregiver to adjust the lighting (e.g., close/open the curtains in the room, or turn on/off a lamp) and the video was played again. To test the audio, the video included music that was presented at the same intensity level as the auditory stimuli in the word-learning task. Caregivers were asked to adjust the volume on their computer if the sound was too loud or not loud enough, until they confirmed that they could hear the music at a comfortable listening level. Once all checks were completed, the experimenters turned off their cameras (so that they would not be visible to the child during the task), provided the link to the study video through the chat function in Zoom, and instructed caregivers to start recording the session locally on their computer using the native video recording application for their operating system (e.g., PhotoBooth for Macs and Camera app for Windows). Recording videos locally avoided lags in the

video that would affect later coding. Instead of using BITTSy for stimulus presentation, the task was displayed in the form of a single video that contained all the trials and attention getters, and different versions of the video were created to preserve the counterbalancing described earlier. Caregivers were asked to set the video to full-screen, hit “play,” and close their eyes for the duration of the video. While completing the tasks, children sat on their caregivers’ lap. Other than these changes, the experimental design was identical to the one described for the in-person group.

Once the experimental video had finished playing, the experimenters turned their cameras back on, and guided caregivers through steps on how to upload the video of the testing sessions that they had just generated using a secure file-transfer link. The experimenters remained on the Zoom call until the video had been successfully uploaded (this usually took 3–5 min).

Data Coding

Participant videos for both the in-person and virtual testing sessions were coded offline on a frame-by-frame basis by two trained coders using Datavyu coding software (Datavyu, 2014). All coding files were checked for reliability across coders, and trials for which there was a discrepancy >0.5 s were re-coded by a third coder. The closest of the two coding files were used for final averaging. For participants in the 29–32 month-old age range, this happened on 4.4% of trials when the task was completed in-person, and on 15.4% of trials when the task was completed virtually. For participants in the 47–50 month-old age range, who primarily completed the task virtually, a third coder was needed on 14.5% of trials.

RESULTS

Feasibility of the Virtual Version of the IPLP

As a first step, we examined how many analyzable trials were collected for children who completed the virtual version of the task, compared to children who had completed the in-person version. We focused on the data from the younger 29–32-month-old group first, given that we had a comparable number of participants who had completed the study in each modality. In order for a trial to be included in the final analyses, participants needed to have looked at one of the objects on the screen for a minimum of 500 ms. As discussed in an in-depth methodological review of the IPLP by Delle Luche et al. (2015), there is a great deal of variability across studies regarding the parameters that have been implemented for data rejection and determining trial inclusion. Many studies do not use or report a minimum looking criteria. However, previous work has established that it takes at least 233 ms for young children to program a saccade and produce looks that are linked to the processing of the stimulus (Zangl et al., 2005; Fernald et al., 2006, 2008). With this in mind, extremely short “looks” might not represent fixations that were intentional or directly linked to the child processing the auditory input that they just heard. While in some previous studies using the IPLP trial inclusion was also restricted to trials in which participants were looking at the attention-getter in the center of the screen at the trial onset, Delle Luche et al. (2015) point out that only about half of the studies rely on this

TABLE 2 | Number of analyzable trials.

| Age group | | In-person | Virtual |
|-----------|-----------------|-----------|---------|
| 29–32 | Baseline trials | 3.9 | 4 |
| | Training trials | 11.7 | 12 |
| | Test trials | 7.8 | 8 |
| 47–50 | Baseline trials | 4 | 3.5 |
| | Training trials | 12 | 10.4 |
| | Test trials | 8 | 6.9 |

practice. Furthermore, the use of this center-fixation criteria is primarily common in studies in which trial-start is triggered by an experimenter that is monitoring child behavior online, but less so in studies when trials are automatically interspaced (Swingley, 2003, 2007; Ramon-Casas et al., 2009). Given that (i) in our study the task was presented as part of a video that contained set durations for the attention-getter in between trials, and (ii) we were unable to trigger trial onsets, we did not apply this rule. As shown in **Table 2**, the number of analyzable trials was comparable for children in both the in-person and virtual modalities. This was true for the 29–32-month-old group as well as the 47–50 month-old group suggesting that the level of engagement with the task was similar across the two age groups that we tested. The same parameters for trial inclusion were applied to both age groups.

We also looked at the attrition rate across in-person and virtual testing sessions. Data from an additional 20 participants were excluded from the in-person group due to technical problems ($n = 1$), side bias ($n = 1$), and fussiness ($n = 18$). This attrition rate is similar to what has been previously reported in other in-person IPLP studies that presented toddlers with a word-learning task (Schmale et al., 2012). Data from an additional seven participants were excluded from the virtual group due to technical problems ($n = 3$), environmental distractors ($n = 1$), not meeting the language exposure requirements ($n = 1$), and fussiness ($n = 2$). Fussiness was defined as inattention to the task and included both children who cried during the study or who refused to sit down and look at the screen. The attrition rate for 47–50 month-olds was comparable to what we observed with the toddlers. Specifically, data from an additional 12 participants were excluded due to technical problems ($n = 6$; all virtual appointments), environmental distractors ($n = 1$; virtual appointment), and fussiness ($n = 5$; 3 in-person and two virtual appointments). We had some initial concerns about being able to maintain young children’s attention through a remote testing procedure, given that we expected there to be less control of the environment, and potentially greater distractors in children’s homes while the task was being completed. Furthermore, we expected to lose a greater amount of data due to technical difficulties during the study (e.g., connectivity problems), and coding problems resulting from a greater variability in the quality of participant videos (due to webcams having different resolutions). To our surprise, the attrition rate was considerably lower for children tested in the virtual group compared to the

in-person group. We found that participants appeared to be more comfortable in their home environment. For example, children tested in the lab more frequently wanted to get up and leave the testing booth, while children in the virtual group were more often content and remained seated in front of the screen for a longer duration. While there were some instances in which a distractor was present in the home and affected task completion for children in the virtual group (e.g., a dog barking, or a sibling talking during the exact time in which the IPLP task was being completed), this was not the norm. Additionally, in the virtual testing procedure, families did not need to travel to the lab, which meant that there was more flexibility to conduct testing sessions in a time-period that aligned better with children's schedules/routines (e.g., testing children right after they had woken up from a nap and were rested). As discussed in our limitations section later on, these parameters might be linked to the demographic characteristics of the sample (e.g., socioeconomic status), making it important to conduct further virtual work with more diverse groups of children.

Differences in Performance Across Testing Modalities

Next, we wanted to evaluate actual performance on the word learning task and compare the data for children who were tested in-person, to that of children who completed the task from home. As a starting point, we examined children's looking time to the objects during the baseline (silent) trials. This was done to ensure there were no pre-existing biases. During these trials children in the in-person group looked at the object on the left on average 50% of the time ($SD = 0.11$) and the object on the right on average 50% of the time ($SD = 0.11$), which is what we would expect since they were not told which object to look at. Similarly, children in the virtual group looked at the object on the left on average 49% of the time ($SD = 0.08$) and the object on the right on average 51% of the time ($SD = 0.08$).

Accuracy was calculated based on the amount of time that the participants remained fixated on the appropriate image, as a proportion of the total time spent fixating on either of the two pictures, averaged over a time window of 300–5100 ms after the onset of the first repetition of the target word, across all test trials of the same condition. This window of analysis was longer than what has been previously used during word recognition tasks (Byers-Heinlein et al., 2017), and this was done given that in the present task children were asked to identify newly-acquired words—rather than highly familiar items (a more difficult task that required additional processing time). Fixating on the appropriate image in this case included the “trained object” on test trials when it was requested, and the “untrained object” on trials when the novel word was requested. This meant that each object was the “correct” item on one of the two test trials but not the other, and if children had learned the target words, they should accurately look at the correct object during both trial types. In fact, two-tailed t -tests indicated that there was no significant difference in accuracy between trained and untrained test trials for the *in-person* modality [$t(18) = 2.04$, $p > 0.05$, Cohen's $d = 0.47$], nor the *virtual* modality [$t(18) = 0.68$,

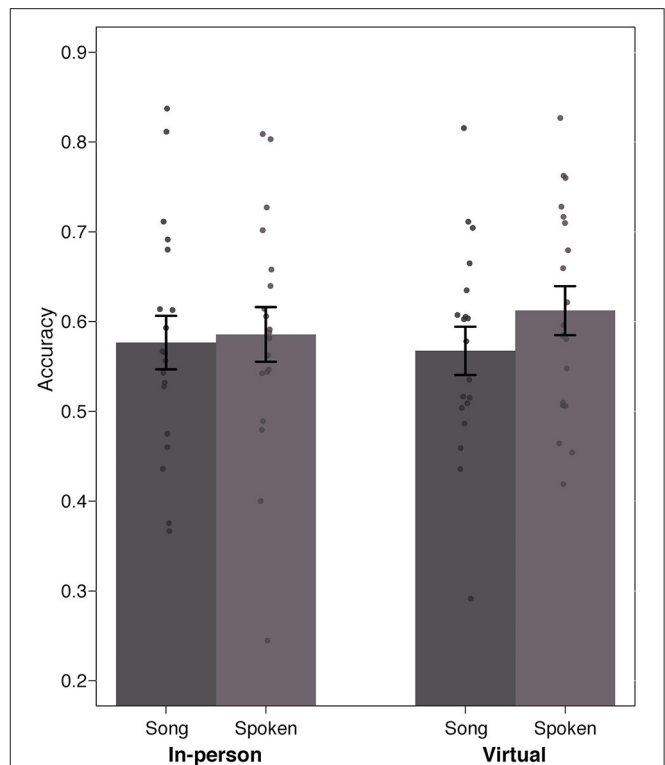


FIGURE 3 | Accuracy data based on proportion of looking to the correct object across the Song and Spoken condition in 29–32 month-olds.

$p > 0.05$, Cohen's $d = 0.16$]. Hence, for the subsequent analyses we collapsed across the two types of test trials.

As shown in **Figure 3**, children's fixation patterns revealed that in general, accuracy was similar in the spoken condition (*in-person* modality: $M = 0.59$, $SD = 0.13$; *virtual* modality: $M = 0.61$, $SD = 0.12$) and in the song condition (*in-person* modality: $M = 0.58$, $SD = 0.13$; *virtual* modality: $M = 0.57$, $SD = 0.11$). A 2×2 mixed ANOVA with Modality as a between-subjects factor (in-person vs. virtual) and Training Condition as a within-subjects factor (spoken vs. song) indicated that there was no significant main effect of training condition [$F_{(1,36)} = 1.69$, $p > 0.05$, $\eta_p^2 = 0.048$] nor modality [$F_{(1,36)} = 0.06$, $p > 0.05$, $\eta_p^2 = 0.001$], and no significant interaction [$F_{(1,36)} = 0.72$, $p > 0.05$, $\eta_p^2 = 0.02$]. This means that (i) the modality in which the study was completed (i.e., in the lab vs. virtually) did not affect children's performance on the task, and (ii) training words in the song condition did not lead to better performance during testing compared to when training occurred in the spoken sentences.

It is also worth noting that two-tailed single-sample t -tests indicated that children across the two modalities performed significantly above chance (in this case 50%) when the training occurred in the spoken condition [*in-person*: $t(18) = 2.81$, $p < 0.05$, Cohen's $d = 0.65$; *virtual*: $t(18) = 4.11$, $p < 0.001$, Cohen's $d = 0.94$], as well as in song [*in-person*: $t(18) = 2.58$, $p < 0.05$, Cohen's $d = 0.59$; *virtual*: $t(18) = 2.51$, $p < 0.05$, Cohen's $d = 0.58$].

The Role of Song on Novel Word Learning in 29–32 Month-Olds

Another goal of the study was to evaluate whether or not using songs during training would facilitate word learning. We found no evidence of this. Our data indicated that 29–32 month-olds successfully acquired novel word-object relations during our task (as indicated by the above-chance performance), but this was equally true when training occurred in a song and in a spoken sentence. One interesting pattern, however, was that the average amount of time that children spent looking at the screen during training trials (arguably a measure of attention) was greater in the song condition than in the spoken condition. This was true for children in both the in-person modality (*song*: $M = 6.8$ s, $SD = 0.64$; *spoken*: $M = 6.3$ s, $SD = 0.69$; $t(18) = 2.81$, $p < 0.05$, two-tailed, Cohen's $d = 0.65$) as well as the virtual modality (*song*: $M = 6.4$ sec, $SD = 0.94$; *spoken*: $M = 5.8$ s, $SD = 1.30$; $t(18) = 2.36$, $p < 0.05$, two-tailed, Cohen's $d = 0.54$). While this pattern of greater “attention” when listening to songs (compared to spoken sentences) aligns with previous research on this topic (Corbeil et al., 2016), our findings would suggest that greater attention (i.e., longer looking times) during training, does not necessarily lead to better learning of the word-object mappings. To our knowledge this is the first study examining the role of song on the acquisition of word-object relations in young children's native language, and it is unclear whether the same pattern of results would extend to other age groups.

The Role of Song on Novel Word Learning in 47–50 Month-Olds and an Examination of Age-Related Differences in Performance

To answer our last research question, we examined whether the testing procedure that we had implemented with toddlers, could also be successfully used with 47–50 month-olds to test their ability to learn novel words in the song and spoken conditions, and whether there were any age-related differences in performance between toddlers and this slightly older group. As a reminder, the majority of participants in the 47–50 month-old group completed the study virtually. Given that we found no significant difference in performance across testing modalities in our previous analyses, we collapsed across the two modalities for the subsequent results.

As an initial step, we examined looking times during baseline trials. We found that 47–50-month-olds looked at the object on the left on average 51% of the time ($SD = 0.11$) and the object on the right on average 49% of the time ($SD = 0.11$), suggesting that there were no pre-existing side biases. Accuracy during test trials was calculated using the same considerations and time window described earlier. Once again, two-tailed t -tests indicated that there was no significant difference in accuracy between trained and untrained test trials [$t(20) = 2.02$, $p > 0.05$, Cohen's $d = 0.44$]; therefore, we collapsed across the two trial types. As shown in **Figure 4**, fixation patterns revealed that surprisingly, accuracy was higher in the spoken condition ($M = 0.69$, $SD = 0.11$) than in the song condition ($M = 0.58$, $SD = 0.15$), and this difference was significant [$t(20) = 2.71$, $p < 0.05$, two-tailed, Cohen's $d = 0.59$]. Additionally, two-tailed single-sample t -tests indicated

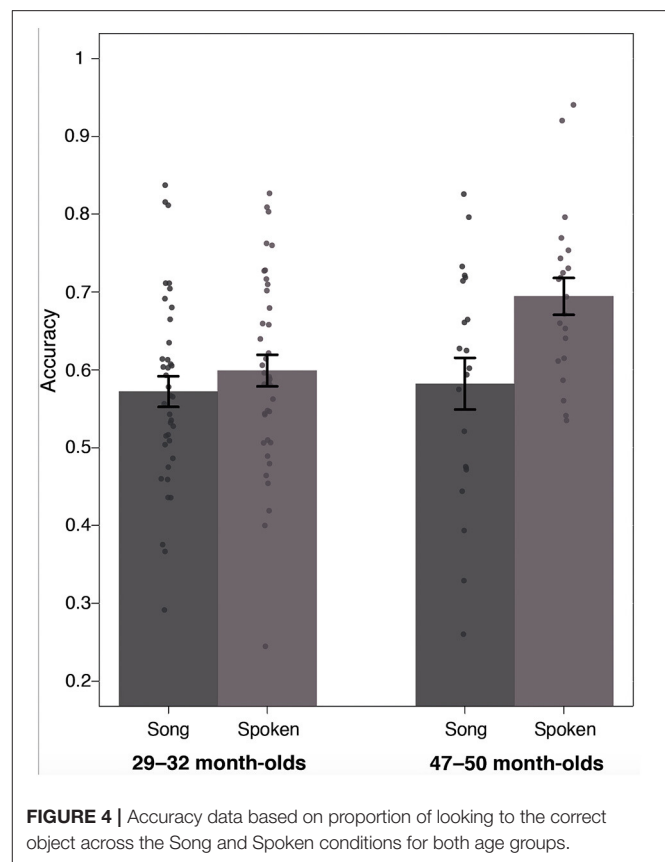


FIGURE 4 | Accuracy data based on proportion of looking to the correct object across the Song and Spoken conditions for both age groups.

that accuracy for the 47–50 month-olds was significantly above chance in both the spoken [$t(20) = 8.22$, $p < 0.001$, Cohen's $d = 1.79$], and the song condition [$t(20) = 2.48$, $p < 0.05$, Cohen's $d = 0.54$], suggesting that children in this age group were also successfully learning the novel word-object pairings.

To examine possible age-related differences, we ran a 2×2 mixed ANOVA with Age as a between-subjects factor (29–32 vs. 47–50) and Training Condition as a within-subjects factor (spoken vs. song). This analysis indicated that there was no significant main effect of age [$F_{(1, 57)} = 3.73$, $p > 0.05$, $\eta_p^2 = 0.04$], but there was a significant main effect of condition [$F_{(1, 57)} = 11.1$, $p < 0.001$, $\eta_p^2 = 0.07$], and a significant interaction [$F_{(1, 57)} = 4.163$, $p < 0.01$, $\eta_p^2 = 0.026$]. To further explore the interaction effect, we conducted simple effects analysis. These demonstrated that when word training occurred in the song, there was no significant difference in performance between the age groups [$F_{(1, 114)} = 0.0883$, $p > 0.05$, $\eta_p^2 = 0.001$]. However, when training occurred in the spoken condition there was a significant difference between the groups [$F_{(1, 114)} = 7.691$, $p < 0.05$, $\eta_p^2 = 0.06$], with 47–50 month-olds showing higher accuracy ($M = 0.69$, $SD = 0.11$) compared to 29–32 month-olds ($M = 0.59$, $SD = 0.12$). Together, these data suggest that using a song to familiarize young children with novel words, does not lead to better learning. In fact, in our current task hearing words in the spoken sentences (during training) led to

higher accuracy during testing in the case of the 47–50 month-olds. Accuracy for the song condition was still significantly above chance, which indicates that hearing words in the song did not prevent participants from acquiring the word-object relations. However, the song did not provide a “boost” in learning, as might have been expected based on the prior attention literature. We also examined whether the average amount of time that children spent looking at the screen during training trials was different for the song compared to the spoken condition (as we had seen for the 29–32 month-old group). However, this was not the case for the 47–50 month-olds [*song*: $M = 6.2$ s, $SD = 0.92$; *spoken*: $M = 6.1$ s, $SD = 0.88$; $t(20) = 0.65$, $p > 0.05$, Cohen’s $d = 0.14$]; that is, although “attention” during training was the same for the two conditions, we still found greater performance during test trials in the spoken condition.

DISCUSSION

The present work set out to investigate the feasibility and precision of a modified version of the Intermodal Preferential Looking Paradigm, which relied on the use of virtual appointments and access to video collected through webcams in participants’ homes. Previous studies using the IPLP have primarily used this measure in a controlled lab setting (Golinkoff et al., 2013); however, due to the global pandemic, many researchers have had to transition to remote testing, in order to keep developmental research activities moving forward. This sudden shift in testing practices has raised questions related to the advantages and disadvantages that come along with collecting data in more natural environments, especially when working with young children who are more easily distracted, and when dealing with fine-grained measures (such as eye-gaze). Our work contrasts data collected through a new virtual version of the IPLP, with data collected through a more established (in-person) version of this paradigm. This is a critical step for advancing developmental research and expanding testing procedures in a sustainable and reliable manner.

The methodological aim outlined above was intertwined with an additional goal to examine the role of song on young children’s vocabulary learning. Previous studies examining the use of music and songs as a tool for teaching language have primarily been conducted with school-aged children in foreign language classrooms (Legg, 2009; Coyle and Gómez Gracia, 2014; Pavia et al., 2019). To our knowledge, no previous studies have directly measured whether songs can be used as a tool to facilitate vocabulary learning (specifically word-object relations) in young children who are acquiring their native language. Furthermore, it is unclear whether there might be developmental changes in how children make use of the information included in the auditory signal (e.g., features of the song), during the word learning process. Our work examined these questions with toddlers and preschoolers using a novel word learning task.

With regards to our methodological goal, data from the younger 29–32 month-old group suggest that there were no differences in performance across participants tested in person and children tested virtually. For both groups, the testing paradigm was identical. The main difference was that one group of toddlers completed the task in a controlled environment

(i.e., a quiet booth in the lab)—using the same equipment across participants (i.e., the same screen, speakers, and video camera), while the other group of toddlers participated from home via a live video call—and used whatever computer screen and camera was available to them. The similarity in performance between groups supports the versatility of the IPLP as a measure that can be used in both lab and remote settings. Based on coding-reliability checks we found that a third coder was more often needed for videos collected with the virtual group, likely due to lower-resolution videos being captured through webcams compared to our in-lab camera, but this only led to an 11% increase in third-coders, which was still manageable. Furthermore, the attrition rate was actually lower for children tested in the virtual group compared to the in-person group, and we argue this was a result of (i) children being more comfortable and hence less fussy in their home environment, and (ii) the virtual testing procedure allowing us to accommodate better to children’s schedules/routines since families no longer had to travel to the lab. We also tested 4-year-olds using the same task, with most participants completing the virtual version of the IPLP. Not only were children in this older group able to complete the task, but coding and attrition rates were comparable to what we had observed with the younger group. Hence, this step allowed us to extend the feasibility of the remote testing approach to slightly older children. It is worth noting that our task only took 7 min to complete, and so the brief duration likely prevented an increase in issues related to children’s attention, and opportunity for distractors to interfere with testing in the home—as might have been the case had the task been longer. It is therefore important to expand this work to other tasks, to examine how different durations and dependent measures might affect the feasibility of collecting data remotely.

Our investigation also provided important insight into the role of song on the acquisition of word-object relations. Children aged 29–32-months were successful at learning novel words, but performance was the same for both words trained in the song condition, as well as in the spoken condition. In other words, we did not find evidence of a facilitatory effect during learning associated with hearing novel words in a song. Children aged 47–50-months once again were accurate in identifying novel word-object pairs that were trained during the task. However, for this older group, performance was higher for words trained in the spoken compared to the song trials. Together, these results suggest that there are age-related differences in how children make use of the auditory information they are presented with while attempting to link words with referents. They also suggest that the use of songs might not facilitate word learning in a native language for toddlers and preschoolers.

These results do not align with (i) previous findings with infants, in which songs were linked to benefits in the acquisition of auditory patterns (Thiessen and Saffran, 2009; Lebedeva and Kuhl, 2010), nor (ii) studies with school-aged children who showed a facilitatory effect of songs when learning a second language (Coyle and Gómez Gracia, 2014). There are some possible explanations for this. First, in the studies with infants, participants simply had to identify sequences of sounds. In the present study, it was necessary to make connections between the auditory patterns (in this case the novel words) and the referents

during training, and subsequently rely on those relations to look at the target object on the screen during testing. Second, in the literature with children who were acquiring an L2, the songs were used across multiple training sessions over a longer period of time (i.e., there were more opportunities to hear the song), and testing was not conducted immediately after a single exposure to the training stimuli (i.e., it was more a measure of retention, rather than immediate recall of the words). This means that the tasks across studies were arguably different and were measuring different abilities. Under this view, it is important to refrain from making overarching conclusions about the role of songs across different types of learning tasks, given that benefits associated with this type of input appear to be task-specific.

There are however, some studies that have reported similar patterns to the ones observed in our data. This comes from tasks in which children were taught content knowledge information in classroom settings. Calvert and Billingsley (1998) examined preschooler's ability to learn their phone number. They found that children were more accurate at remembering their phone number when it was presented to them in speech rather than song. In that same paper, they also discussed data indicating that while repeated exposure to a song improved verbatim word-for-word memory of lyrics in an unfamiliar language (in this case incomprehensible French), it did not facilitate recall of words in a familiar language. Similar findings were identified in a study with second-grade students in which information about historical events was trained either in songs or in speech, and later assessed (Calvert, 2001). Once again, songs led to improvement in verbatim memory, but only training in the spoken condition was associated with better retention of content knowledge. The authors propose that there are different "levels of learning," from more superficial processing of information (e.g., verbatim word-for-word memory, in which the actual meaning is not retained), to deeper learning (e.g., encoding and retrieving the details about the historical events). Furthermore, songs might be more conducive to superficial-level learning, as children may focus on superficial qualities of song (e.g., the rhyming, melody) rather than the content information.

This theoretical explanation could help us understand why preschoolers in our study had higher accuracy in the spoken condition compared to the song condition. Our task was challenging, as it required participants to understand the relation between the objects and the words to accurately look at the target object during trained test trials. In addition, children had to use that information along with their understanding of mutual exclusivity to also look at the correct object during untrained test trials. These steps likely required a deeper level of learning than if children were simply tested on their ability to recognize that they had heard the word "doop" based on verbatim memory, without knowing its meaning (i.e., what referent it corresponded to). In the case of the 29–32 month-old group, overall performance in the task was lower compared to the older participants, so it is possible that the task was simply more challenging for the younger group. In other words, given the difficulty of the task, it may not have been sensitive enough to capture differences that may exist between the use of speech

and song for learning word-object relations in toddlers. We acknowledge this as a limitation of the study.

There are other elements that may have limited our findings. First, the modality of the testing trials required participants to generalize words across song and speech. As a reminder, in our paradigm children were trained in either spoken sentences or in a song (depending on the block), but all testing trials were presented in spoken sentences. This meant that in the song blocks, children had to recognize that the word "doop" that was sung during training, was the same word "doop" that was spoken during testing. We chose this methodological approach because it is one that has been used in previous studies with young children (Thiessen and Saffran, 2009). Additionally, given that in the real world children must rely on spoken sentences for oral communication and social interactions, this type of generalization is critical if songs are to be used as a way of supporting language learning. We do know, however, that infants have difficulty identifying words that they heard during familiarization when there were differences in the speech signal during testing; for example, hearing a word in a happy voice and later hearing it in a neutral or sad voice (Singh, 2008). Given that song exaggerates features of speech, there may have been a similar disadvantage at play, when children had to generalize from song to speech in our study. To examine this possibility, future work should manipulate the modality of the testing trials to see if a change that eliminates the need to generalize words in the song condition would lead to a different pattern of performance.

A second point related to the characteristics of the speech stimuli, is that sentences in the spoken condition were produced using infant-directed speech prosody. As stated in the introduction, IDS contains melodic features that make it more similar to songs compared to say adult-directed speech (ADS). The methodological decision to use IDS was made given that previous studies that used the IPLP to examine word learning in toddlers have used this type of speech register (Schmale et al., 2011; Newman et al., 2018), and because IDS has been found to increase attention and guide word learning in toddlers (Nencheva et al., 2021). Nevertheless, it is possible that adding a condition in which spoken sentences are produced in ADS might lead to even better accuracy during this type of learning task, and perhaps even lead to a difference in performance with the younger participants. This step would offer a good comparison since the spoken sentences would be less melodic and more distinct from the song condition, and would provide a better understanding of what might be driving the effects that were observed with the present data.

Third, in our study, children were only presented with a limited number of training trials, and testing was only carried out immediately after training. While this is a type of design that has been previously used in word-learning studies with children of similar ages (Schmale et al., 2012; Newman et al., 2018), it limits our ability to examine whether variations in the amount of training may lead to songs providing a benefit. For example, in real-world scenarios, children have more than three exposures to a novel word-object pair. Furthermore, we only tested children on their ability to identify words immediately after being familiarized with the novel words. It is possible that additional testing that

is delayed (e.g., a week after training) might provide information about the retention of information that children learned during the task, and whether songs and spoken sentences affect retention of the words differently. These questions should be explored in future work.

Fourth, the use of a familiar melody in the song condition may have posed an additional challenge. The study used the tune of “Old MacDonald had a Farm”—changing only the words of the song. Using familiar melodies and changing the lyrics to introduce new concepts is a common practice in educational settings with children of different ages (Wolfe and Hom, 1993). However, it is possible that the use of a familiar melody during training may have resulted in some level of confusion, as children could have been anticipating the familiar lyrics rather than those presented to them. Based on parent report, 100% of the children in the 47–50-month group were familiar with the song “Old MacDonald had a Farm,” as were 100% of the children in the 29–32-month virtual group. Additionally, 16 of the 19 children in the 29–32-month in-person group were familiar with the song, and parents of the remaining three children were unsure if their children knew the song. This meant that the vast majority of participants who completed the task knew the song and may have anticipated hearing the “traditional” words. While performance in the song condition was still above chance for both age groups—suggesting that the song was not preventing children from learning the word-object relations altogether—a potential boost in learning from the song may have been hampered by pre-existing expectations about the melody. An interesting follow-up study would be to use an unfamiliar melody during the training phase, as this would remove prior experience with the song lyrics.

Lastly, there are limitations associated with the demographic characteristics of the children that were included in the present work. It is important to first note that our sample included primarily children from households with mid-to-high socioeconomic status (SES). This was true for both age groups. Additionally, to participate in the virtual version of the study, families were required to have access to high-speed internet and a computer with a webcam, which limited participation opportunities for some families. Nevertheless, barriers exist for in-person studies as well. In many cases, families must have access to transportation, as well as available time during lab operating hours to visit the lab and complete the testing session. Some ways to mitigate the in-person obstacles have been to provide funds for transportation and to offer flexible testing hours. There are also potential ways of addressing barriers associated with online testing that are worth considering, which include providing families with hot spots for internet access, and offering temporary access to technological devices (e.g., loaner computers). A critical next step is therefore, to extend this work to more diverse groups of children, as it will improve our ability to generalize the results.

To conclude, findings from the present study support the feasibility of using a virtual version of the IPLP to collect

remote eye-gaze data in both toddlers and preschool children. This serves as a way of continuing to move forward with developmental language research, during situations when it is not possible for in-person interactions with participants to take place. Additionally, we provide evidence suggesting that using songs during vocabulary training does not result in better learning, and that providing linguistic information to young children through spoken sentences might lead to improved outcomes. These findings hold implications not only for learning-through-song interventions, but also for instruction in educational settings. While using songs may help increase attention during a particular task, greater attention may not equate to deep-level learning. Therefore, using songs may help increase engagement (and perhaps participation), but when introducing new concepts for children to retain, using spoken sentences may be best.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found at: <https://osf.io/qmb6t/>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board at the University of Delaware. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was funded by start-up funds to GM from the University of Delaware, and by a Unidel Foundation fellowship to MB.

ACKNOWLEDGMENTS

We thank Katherine Richard, Madison Pruitt, Sarah Dombroski, Shakhlo Nematova, Abhigna Rao, Aurora Reible-Gunter, Lauren Mellor, Ben Cushman, Claudia Kurtz, Sarah Blum, Aashaka Desai, Kathryn Catalino, Katrina Conner, Mariann Angela Agapito, Shruti Shirapu, Paige Kassman, Jillian Lardiere, Talia Gillespie, Silpa Annavarapu, Jessica Price, Brianna Postorino, Dea Harjianto, Sydney Horne, Ryan Moore, Taylor Hallacy, Emily Fritzson, Kathryn Catalino, Nicole Khanutin, Jackson Xiao, Chaithra Reddy, Emily Arena, Nicole Scacco, Sophia Emery, Elizabeth Smith, Erin Felter, Kathryn Hall, Katherine Filliben, Amanda Kalil, and Hannah Wissner for assistance in coding, scheduling, and testing participants.

REFERENCES

- Bergelson, E., and Swingle, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proc. Natl. Acad. Sci. U.S.A.* 109, 3253–3258. doi: 10.1073/pnas.111380109
- Brandt, A., Gebrian, M., and Slevc, L. R. (2012). Music and early language acquisition. *Front. Psychol.* 3:327. doi: 10.3389/fpsyg.2012.00327
- Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.
- Busse, V., Jungclaus, J., Roden, I., Russo, F. A., and Kreutz, G. (2018). Combining song-and speech-based language teaching: an intervention with recently migrated children. *Front. Psychol.* 9:2386. doi: 10.3389/fpsyg.2018.02386
- Byers-Heinlein, K., Morin-Lessard, E., and Lew-Williams, C. (2017). Bilingual infants control their languages as they listen. *Proc. Natl. Acad. Sci.* 114, 9032–9037. doi: 10.1073/pnas.1703220114
- Calvert, S. L. (2001). Impact of televised songs on children's and young adults' memory of educational content. *Media Psychol.* 3, 325–342. doi: 10.1207/S1532785XMEP0304_02
- Calvert, S. L., and Billingsley, R. L. (1998). Young children's recitation and comprehension of information presented by songs. *J. Appl. Dev. Psychol.* 19, 97–108. doi: 10.1016/S0193-3973(99)80030-6
- Capone, N. C., and McGregor, K. K. (2005). The effect of semantic representation on toddlers' word retrieval. *J. Speech Lang. Hear. Res.* 48, 1468–1480. doi: 10.1044/1092-4388(2005)102
- Certain, L. K., and Kahn, R. S. (2002). Prevalence, correlates, and trajectory of television viewing among infants and toddlers. *Pediatrics* 109, 634–642. doi: 10.1542/peds.109.4.634
- Corbeil, M., Trehub, S. E., and Peretz, I. (2013). Speech vs. singing: Infants choose happier sounds. *Front. Psychol.* 4:372. doi: 10.3389/fpsyg.2013.00372
- Corbeil, M., Trehub, S. E., and Peretz, I. (2016). Singing delays the onset of infant distress. *Infancy* 21, 373–391. doi: 10.1111/inf.12114
- Coyle, Y., and Gómez Gracia, R. (2014). Using songs to enhance L2 vocabulary acquisition in preschool children. *ELT J.* 68, 276–285. doi: 10.1093/elt/ccu015
- Cross, I., and Morley, I. (2008). "The evolution of music: theories, definitions, and the nature of the evidence," in *Communicative Musicality*, eds S. N. Malloch and C. Trevarthen (Oxford: Oxford University Press), 61–82.
- Datavyu, T. (2014). *Datavyu: A Video Coding Tool*. In *Databrary Project*. New York, NY: New York University. Available online at: <http://datavyu.org>
- Delle Luche, C., Durrant, S., Poltrok, S., and Floccia, C. (2015). A methodological investigation of the Intermodal Preferential Looking paradigm: methods of analyses, picture selection, and data rejection criteria. *Infant Behav. Dev.* 40, 151–172. doi: 10.1016/j.infbeh.2015.05.005
- Fenson, L., Dale, P., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., et al. (1994). Variability in early communicative development. *Monogr. Soc. Res. Child Dev.* 59, 1–173. doi: 10.2307/1166093
- Fernald, A. (1992). "Meaningful melodies in mothers' speech to infants," in *Nonverbal Vocal Communication: Comparative and Developmental Approaches*, eds H. Papoušek and U. Jürgens (https://www.google.com/search?xsrf=ALeKk020Baqe1mAy0bzAoX4JH05RD7ZJQ:1626713089081&q=Cambridge&stick=H4sIAAAAAAAAAOPgE-LQz9U3yCowyFICsyZNLyq0tLKTrfTzi9IT8zKrEksy8_NQOFYZqYkphaWJRSWpRcWLWDmdE3OTijT0lN3sDICAPVRwPRAAAA&sa=X&ved=2ahUKEwiNuM2-yu_xAhWbzzgGHdY3DMAQmxMoATAZegQINRAD Cambridge: Cambridge University Press), 262–282.
- Fernald, A., Perfors, A., and Marchman, V. A. (2006). Picking up speed in understanding: speech processing efficiency and vocabulary growth across the 2nd year. *Dev. Psychol.* 42, 98–116. doi: 10.1037/0012-1649.42.1.98
- Fernald, A., and Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Dev. Psychol.* 20, 104–113. doi: 10.1037/0012-1649.20.1.104
- Fernald, A., Swingle, D., and Pinto, J. P. (2001). When half a word is enough: infants can recognize words using partial phonetic information. *Child Dev.* 72, 1003–1015. doi: 10.1111/1467-8624.00331
- Fernald, A., Zangl, R., Portillo, A. L., and Marchman, V. A. (2008). Looking while listening: using eye movements to monitor spoken language. *Dev. Psycholinguist. On-line Methods Child. Lang. Process.* 44:97. doi: 10.1075/lald.44.06fer
- Frank, M. C., Alcock, K. J., Arias-Trejo, N., Aschersleben, G., Baldwin, D., Barbu, S., et al. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* 3, 24–52. doi: 10.1177/2515245919900809
- Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., and Gordon, L. (1987). The eyes have it: lexical and syntactic comprehension in a new paradigm. *J. Child Lang.* 14, 23–45. doi: 10.1017/S030500090001271X
- Golinkoff, R. M., Ma, W., Song, L., and Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: what have we learned? *Perspect. Psychol. Sci.* 8, 316–339. doi: 10.1177/1745691613484936
- Good, A. J., Russo, F. A., and Sullivan, J. (2015). The efficacy of singing in foreign-language learning. *Psychol. Music* 43, 627–640. doi: 10.1177/0305735614528833
- Gupta, P. (2005). What's in a word? a functional analysis of word learning. *Perspect. Lang. Learn. Educ.* 12, 4–8. doi: 10.1044/llc12.3.4
- Halberda, J. (2003). The development of a word learning strategy. *Cognition* 87, B23–B34. doi: 10.1016/S0010-0277(02)00186-5
- Hollich, G., Hirsh-Pasek, K., and Golinkoff, R. M. (2000). Breaking the language barrier: an emergentist coalition model of word learning. *Monogr. Soc. Res. Child Dev.* 65, 1–123. doi: 10.1111/1540-5834.00090
- Houston-Price, C., Mather, E., and Sakkalou, E. (2007). Discrepancy between parental reports of infants' receptive vocabulary and infants' behaviour in a preferential looking task. *J. Child Lang.* 34, 701–724. doi: 10.1017/S0305000907008124
- Jackendoff, R. (2009). Parallels and nonparallels between language and music. *Music Percept.* 26, 195–204. doi: 10.1525/mp.2009.26.3.195
- Lebedeva, G. C., and Kuhl, P. K. (2010). Sing that tune: infants' perception of melody and lyrics and the facilitation of phonetic recognition in songs. *Infant Behav. Dev.* 33, 419–430. doi: 10.1016/j.infbeh.2010.04.006
- Legg, R. (2009). Using music to accelerate language learning: an experimental study. *Res. Educ.* 82, 1–12. doi: 10.7227/RIE.82.1
- Ludke, K. M., Ferreira, F., and Overy, K. (2014). Singing can facilitate foreign language learning. *Mem. Cogn.* 42, 41–52. doi: 10.3758/s13421-013-0342-5
- Ma, W., Golinkoff, R. M., Houston, D. M., and Hirsh-Pasek, K. (2011). Word learning in infant-and adult-directed speech. *Lang. Learn. Dev.* 7, 185–201. doi: 10.1080/15475441.2011.579839
- McDowd, J. M. (2007). An overview of attention: behavior and brain. *J. Neurol. Phys. Ther.* 31, 98–103. doi: 10.1097/NPT.0b013e31814d7874
- Merriman, W. E., and Bowman, L. L. (1989). The mutual exclusivity bias in children's word learning. *Monogr. Soc. Res. Child Dev.* 54, 1–129. doi: 10.2307/1166130
- Morini, G., and Newman, R. S. (2019). Dónde está la ball? examining the effect of code switching on bilingual children's word recognition. *J. Child Lang.* 46, 1238–1248. doi: 10.1017/S0305000919000400
- Nakata, T., and Trehub, S. E. (2004). Infants' responsiveness to maternal speech and singing. *Infant Behav. Dev.* 27, 455–464. doi: 10.1016/j.infbeh.2004.03.002
- Nencheva, M. L., Piazza, E. A., and Lew-Williams, C. (2021). The moment-to-moment pitch dynamics of child-directed speech shape toddlers' attention and learning. *Dev. Sci.* 24:e12997. doi: 10.1111/desc.12997
- Newman, R. S., Morini, G., Kozlovsky, P., and Panza, S. (2018). Foreign accent and toddlers' word learning: the effect of phonological contrast. *Lang. Learn. Dev.* 14, 97–112. doi: 10.1080/15475441.2017.1412831
- Newman, R. S., Shroads, E. A., Johnson, E. K., Kamdar, J., Morini, G., Onishi, K., et al. (2021). Introducing BITTSy: the behavioral infant and toddler testing system. *Behav. Res. Methods* doi: 10.3758/s13428-021-01583-9. [Epub ahead of print].
- Overland, C. T. (2017). Music education, Inc. *Music Educ. J.* 104, 55–61. doi: 10.1177/0027432117719462
- Pavia, N., Webb, S., and Faez, F. (2019). Incidental vocabulary learning through listening to songs. *Stud. Second Lang. Acquis.* 41, 745–768. doi: 10.1017/S0272263119000020
- Peretz, I. (2009). Music, language, and modularity framed in action. *Psychol. Belg.* 49, 157–175. doi: 10.5334/pb-49-2-3-157
- Pinker, S. (1997). *How the Mind Works*. Manhattan, NY: Norton.
- Ramon-Casas, M., Swingle, D., Sebastián-Gallés, N., and Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cogn. Psychol.* 59, 96–121. doi: 10.1016/j.cogpsych.2009.02.002

- Salcedo, C. S. (2010). The effects of songs in the foreign language classroom on text recall, delayed text recall and involuntary mental rehearsal. *J. Coll. Teach. Learn.* 7, 19–30. doi: 10.19030/tlc.v7i6.126
- Schmale, R., Cristia, A., and Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Dev. Sci.* 15, 732–738. doi: 10.1111/j.1467-7687.2012.01175.x
- Schmale, R., Hollich, G., and Seidl, A. (2011). Contending with foreign accent in early word learning. *J. Child Lang.* 38, 1096–1108. doi: 10.1017/S0305000910000619
- Shatz, M. (1978). Children's comprehension of their mothers' question-directives. *J. Child Lang.* 5, 39–46. doi: 10.1017/S0305000900001926
- Simpson, K., and Keen, D. (2009). Teaching young children with autism graphic symbols embedded within an interactive song. *J. Dev. Phys. Disabil.* 22, 165–177. doi: 10.1007/s10882-009-9173-5
- Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition* 106, 833–870. doi: 10.1016/j.cognition.2007.05.002
- Singh, L., Nestor, S., Parikh, C., and Yull, A. (2009). Influences of infant-directed speech on early word recognition. *Infancy* 14, 654–666. doi: 10.1080/15250000903263973
- Snijders, T. M., Benders, T., and Fikkert, P. (2020). Infants segment words from songs—an EEG study. *Brain Sci.* 10:39. doi: 10.3390/brainsci10010039
- Stager, C. L., and Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature* 388, 381–382. doi: 10.1038/41102
- Stern, D. N., Spieker, S., Barnett, R. K., and Mackain, K. (1983). The prosody of maternal speech: infant age and context related changes. *J. Child Lang.* 10, 1–15. doi: 10.1017/S0305000900005092
- Stern, D. N., Spieker, S., and Mackain, K. (1982). Intonation contours as signals in maternal speech to prelinguistic infants. *Dev. Psychol.* 18, 727–735. doi: 10.1037/0012-1649.18.5.727
- Swingle, D. (2003). Phonetic detail in the developing lexicon. *Lang. Speech* 46, 265–294. doi: 10.1177/00238309030460021001
- Swingle, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Dev. Psychol.* 43, 454–464. doi: 10.1037/0012-1649.43.2.454
- Swingle, D., and Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychol. Sci.* 13, 480–484. doi: 10.1111/1467-9280.00485
- Thiessen, E. D., Hill, E. A., and Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy* 7, 53–71. doi: 10.1207/s15327078in0701_5
- Thiessen, E. D., and Saffran, J. R. (2009). How the melody facilitates the message and vice versa in infant learning and memory. *Ann. N. Y. Acad. Sci.* 1169, 225–233. doi: 10.1111/j.1749-6632.2009.04547.x
- Tincoff, R., and Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychol. Sci.* 10, 172–175. doi: 10.1111/1467-9280.00127
- Trainor, L. J. (1996). Infant preferences for infant-directed versus noninfant-directed playsongs and lullabies. *Infant Behav. Dev.* 19, 83–92. doi: 10.1016/S0163-6383(96)90046-6
- Trainor, L. J., Clark, E. D., Huntley, A., and Adams, B. A. (1997). The acoustic basis of preferences for infant-directed singing. *Infant Behav. Dev.* 20, 383–396. doi: 10.1016/S0163-6383(97)90009-6
- Trehub, S. E. (2003). The developmental origins of musicality. *Nat. Neurosci.* 6, 669–673. doi: 10.1038/nn1084
- Trehub, S. E., Hill, D. S., and Kamenetsky, S. B. (1997a). Parents' sung performances for infants. *Can. J. Exp. Psychol.* 51, 385–396. doi: 10.1037/1196-1961.51.4.385
- Trehub, S. E., and Trainor, L. (1993). “Listening strategies in infancy: the roots of language and musical development,” in *Cognitive Aspects of Human Audition*, eds S. McAdams and E. Bigand (Oxford: Oxford University Press), 278–327.
- Trehub, S. E., and Trainor, L. J. (1998). Singing to infants: lullabies and play songs. *Adv. Infancy Res.* 12, 43–78.
- Trehub, S. E., Unyk, A. M., Kamenetsky, S. B., Hill, D. S., Trainor, L. J., and Henderson, J. L. (1997b). Mothers' and fathers' singing to infants. *Dev. Psychol.* 33, 500–507. doi: 10.1037/0012-1649.33.3.500
- Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., and Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Dev. Psychol.* 34:1289. doi: 10.1037/0012-1649.34.6.1289
- Werker, J. F., Fennell, C. T., Corcoran, K. M., and Stager, C. L. (2002). Infants' ability to learn phonetically similar words: effects of age and vocabulary size. *Infancy* 3, 1–30. doi: 10.1207/S15327078IN0301_1
- Wolfe, D. E., and Hom, C. (1993). Use of melodies as structural prompts for learning and retention of sequential verbal information by preschool students. *J. Music Ther.* 30, 100–118. doi: 10.1093/jmt/30.2.100
- Wolfe, D. E., and Noguchi, L. K. (2009). The use of music with young children to improve sustained attention during a vigilance task in the presence of auditory distractions. *J. Music Ther.* 46, 69–82. doi: 10.1093/jmt/46.1.69
- Zangl, R., Klarman, L., Thal, D., Fernald, A., and Bates, E. (2005). Dynamics of word comprehension in infancy: developments in timing, accuracy, and resistance to acoustic degradation. *J. Cogn. Dev.* 6, 179–208. doi: 10.1207/s15327647jcd0602_2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Morini and Blair. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Web-Based Auditory and Visual Emotion Perception Task Experiment With Children and a Comparison of Lab Data and Web Data

Hisako W. Yamamoto^{1,2}, Misako Kawahara^{1,2} and Akihiro Tanaka^{1*}

¹ Tokyo Woman's Christian University, Tokyo, Japan, ² Japan Society for the Promotion of Science, Tokyo, Japan

OPEN ACCESS

Edited by:

Sho Tsuji,
The University of Tokyo, Japan

Reviewed by:

Yang Yang,
Nanyang Technological University,
Singapore
Paddy Ross,
Durham University, United Kingdom

*Correspondence:

Akihiro Tanaka
akih.tanaka@gmail.com

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 29 April 2021

Accepted: 19 July 2021

Published: 18 August 2021

Citation:

Yamamoto HW, Kawahara M and
Tanaka A (2021) A Web-Based
Auditory and Visual Emotion
Perception Task Experiment With
Children and a Comparison of Lab
Data and Web Data.
Front. Psychol. 12:702106.
doi: 10.3389/fpsyg.2021.702106

Due to the COVID-19 pandemic, the significance of online research has been rising in the field of psychology. However, online experiments with child participants are rare compared to those with adults. In this study, we investigated the validity of web-based experiments with child participants 4–12 years old and adult participants. They performed simple emotional perception tasks in an experiment designed and conducted on the Gorilla Experiment Builder platform. After short communication with each participant via Zoom videoconferencing software, participants performed the auditory task (judging emotion from vocal expression) and the visual task (judging emotion from facial expression). The data collected were compared with data collected in our previous similar laboratory experiment, and similar tendencies were found. For the auditory task in particular, we replicated differences in accuracy perceiving vocal expressions between age groups and also found the same native language advantage. Furthermore, we discuss the possibility of using online cognitive studies for future developmental studies.

Keywords: online experiments, emotion perception, cognitive development, auditory perception, visual perception, vocal expression, facial expression

INTRODUCTION

The COVID-19 pandemic that began in 2020 has forced people to move much of their daily, face-to-face communication online. Psychological experiments are no exception. Many behavioral scientists had to stop their research and decide whether to postpone it or to move it online. Although many researchers have been trying to conduct studies remotely, sufficient examination of the validity of online developmental research is absent to date. In the present study, we introduce an online trial of perception tasks for children. We conducted a simple experiment featuring an auditory and a visual emotion perception task using video chat and an online experiment platform with children (4–12 years old) and adult participants. We then examined the validity of these data (online data) with the data from our previous, similar laboratory trial (Kawahara et al., 2021).

Even before the pandemic, online experiment research targeting adults was becoming popular due to its advantages. Unlike laboratory experiments, in which participants tend to be limited to residents around universities (e.g., Henrich et al., 2010), in online studies researchers can recruit participants without geographical constraints. Moreover, online experiments pair well with crowdsourcing services. Such services enable researchers to collect large amounts of data at low costs within a short time (Stewart et al., 2017). Such advantages have led many cognitive psychology

researchers to adopt data collected through online experiments (e.g., Mills and D'Mello, 2014; Shin and Ma, 2016; Laeng et al., 2018; Lavan et al., 2018; McPherson and McDermott, 2018; Carbon, 2020).

However, can online experiments ensure the validity and quality of the data they generate? To answer this question, some studies have compared online cognitive experiment data with laboratory experiment data and reported their success in replicating results (e.g., Crump et al., 2013; Simcox and Fiez, 2014; de Leeuw and Motz, 2016). Previous studies have also demonstrated some disadvantages of online studies. One such problem is high dropout rates (Reips, 2002; Zhou and Fishbach, 2016). Moreover, even when participants remained until the end of the experiment, some of them, known as “satisfiers,” might not devote the cognitive effort in the tasks (Miura and Kobayashi, 2016). In addition to considering issues with online participants, we should consider the variety of their environments. In most of online studies, participants use their own devices. For this reason, while the validity of online data has been ensured for within-subject designs (e.g., Semmelmann and Weigelt, 2017), the case for between-subject design has not been clearly made. These factors may lead to greater variance in web experiments compared to lab experiments (Germine et al., 2012). Thus, while Internet-based experiments are easy to participate in, there may be some problems due to this ease (see Paolacci and Chandler, 2014).

Now then, what is the situation with online research for child participants? We were able to find some trials and projects that shifted developmental research online. For example, Tran et al. (2017) tried to move an infant study online by recruiting participants through Amazon Mechanical Turk. They measured the length of time that 5–8-month-olds remained looking at various stimuli and reported success in capturing changes in their attention depending on the stimulus presented, even in an online data collection environment. Concerning behavioral measures, Klindt et al. (2017) reported that the large amount of data they collected from online participants revealed changes in cognitive skills (e.g., working memory, false belief, etc.) over the human lifespan. They collected the data through the BRAiN²US online platform for smartphones, and participants also included children (participants ranged from 5 to 85 years old). However, their study did not focus on the validity of online experiments with children, nor did it compare their data with lab data; rather their aim was to obtain a large dataset from a wide range of participants. More recently, Nussenbaum et al. (2020) investigated the decision-making strategies of participants aged 8–25 during an online task. They compared their results with data from previous lab experiments and were able to replicate age-related changes in strategy even in the online experiment. Moreover, some new online platforms for child research, such as Lookit (Scott et al., 2017), Discoveries Online (Rhodes et al., 2020), and Childlab (Sheskin and Keil, 2018) have been developed with the aim of enabling participation in remote studies for children who are not able to easily travel to a laboratory.

These studies notwithstanding, less remote developmental research is being conducted than remote cognitive studies targeting adults. Why have developmental researchers hesitated

to choose online research to pursue their research questions? The lack of online developmental research may be caused by the following difficulties. First, it is difficult for participants to form a rapport with the experimenter during online experiments. A rapport is important for ensuring that child participants are as relaxed as possible while engaging in tasks. Second, we cannot always check whether a participant is really a child (and not an adult), and participant age is a critical factor in developmental research. Third, differences in performance between different aged participants may be difficult to observe because experiments with a between-subject design are not considered suitable for online research. However, as the pandemic continues, the benefits of online developmental research may surpass such disadvantages if we can ensure the data is valid. Thus, it is imperative to accumulate data from online developmental studies featuring various online tasks to determine its suitability for use in future research.

In this study, we report on our attempt at moving an experiment involving children's perception tasks online. Our experiment consisted of video chat communication, and main tasks were controlled through the online experiment platforms. First, the experimenter communicated with the child participants and their parents via Zoom¹ to check the child's participation and to build rapport with them. Next, the experimenter guided participants to the browser experiment webpage built with Gorilla Experiment Builder^{2,3} (Anwyl-Irvine et al., 2020) and instructed them to engage in two simple emotion perception tasks. To investigate the validity of the obtained data, we compared web data for each task lab data for very similar tasks (Kawahara et al., 2021). We hypothesized that this online experiment method would reduce the issues usually associated with an online experiment. Specifically, we predicted that participants would perform as well in the online experiment as in the lab experiment and that the accuracy of each task would not differ between web and lab data.

In the emotion perception tasks, participants were asked to judge emotions by watching dynamic facial expressions or listening to vocal expressions. We chose these tasks for our online developmental research for two reasons. First, the development of emotional perception has not been investigated in online research. Second, the emotional judgment task enables us to examine the effect of stimulus presented through a web browser on auditory (vocal expression) perception and visual (facial expression) perception independently. To compare web data with lab data for each modality, participants engaged in an auditory task judging emotions by listening to sounds only, and a visual task judging emotions by watching facial dynamics only (with no sound).

¹<https://zoom.us/>

²<https://gorilla.sc/>

³Anwyl-Irvine et al. (2020) showed that child participants had completed a flanker task created with Gorilla Experiment Builder (the youngest participant in the final sample was 4.38 years old) and revealed the development of the children's performance. This was not a remote online study because participation in the experiment took place in a laboratory setting with an experimenter, not in participants' homes. Nonetheless, the results do suggest that even child participants can engage in cognitive tasks controlled by this platform.

MATERIALS AND METHODS

Participants

Web Data

The 36 children aged 4–12 years old (30 girls and 6 boys) and the 16 undergraduate or graduate students (age range: 18–29, M age = 21.63, 13 women and 3 men) participated in the experiment. Since one 5-years-old girl's parent reported that she used built-in laptop speakers because her earphones did not fit her, her data for both tasks were excluded from the analysis. In the analysis, 4–8-year-old children were classified as the younger child group ($N = 21$, M age = 6.48 years old) and the 9–12-year-old children were classified as the older child group ($N = 14$, M age = 10.29 years old). Data were collected from undergraduate or graduate students to compare the data collected from children with data collected from adults.

Child participant data were collected during the online science event of the National Museum of Emerging Science and Innovation (Miraikan) in Tokyo, Japan. We recruited participants through the Miraikan web page and SNS services (Twitter, Facebook). This event was held from August to December 2020. Adult participants were recruited through a snowball-sampling method and the Crowdfunder crowdsourcing service website.⁴

All participants spoke Japanese as their native language. All adult participants and parents of child participants were informed of the purpose of the study and gave informed consent in accordance with the Declaration of Helsinki by checking a box on the consent page during the browser experiment session.

Lab Data

We compared lab data from the unimodal session of our previous experiment (Kawahara et al., 2021) with our web data. Data collected from 179 children aged 5–12 years old (75 girls and 104 boys) and from 33 undergraduate or graduate students (age range: 18–32, M age = 22.39, 17 women, 16 men) were included in the analysis. Child participants' lab data were collected during the science event held at the Miraikan in 2015. We recruited participants through the Miraikan web page. For data analysis, the 5–8 year-old children comprised the younger child group ($N = 100$, M age = 6.36), and the 9–12 year-old children comprised the older child group ($N = 79$, M age = 10.66). Adult participants were recruited using a snowball-sampling method. As with participants in the web experiment, all participants in the lab experiment spoke Japanese as their native language. Adult participants and parents of child participants gave written informed consent in accordance with the Declaration of Helsinki.

Stimuli

Web Data

The auditory and visual stimuli used were based on the audiovisual stimuli originally used by Tanaka et al. (2010). These audiovisual stimuli (stimuli used as the “congruent condition” in their study) were short video clips featuring a speaker expressing anger or happiness through face and

voice expression. The speakers were four women (two native Japanese speakers and two native Dutch speakers). In each video clip, each speaker speaks one of four utterances containing only emotionally neutral linguistic information, including Hello (Japanese, “*Hai, moshimoshi*”; Dutch, “*Hallo, dat ben ja*”) and Good-by (Japanese, “*Sayonara*”; Dutch “*Een goede dag*”); What is this? (Japanese, “*Korenani*”; Dutch “*Hey, wat is dit?*”); and Is that so? (Japanese, “*Sounandesuka*”; Dutch, “*Oh, is dat zo?*”). A total of 32 video clips [in two languages (Japanese and Dutch) \times two emotions (angry and happy) \times two speakers \times four utterances] were used.

Auditory stimuli were created by turning off the images and adding a gray rectangle image of the same size. Visual stimuli were created by muting sounds. Auditory stimuli comprised 32 video clips with vocal expression information only. Visual stimuli comprised 32 video clips with facial expression information only. The resolution of each video clip was 640 \times 480 pixels. In the web experiment, auditory and visual stimuli were encoded in MP4 files for web page presentation.

Lab Data

The web experiment stimuli and the lab experiment stimuli were almost same but differed in file format. In the lab experiment, the auditory stimuli files were in WAV format and the visual stimuli files in AVI format. Moreover, in the lab experiment auditory stimuli were presented with a blank, white display, and in web experiment a gray rectangle was displayed while the auditory stimuli were presented. The latter was to prevent web participants from becoming anxious due to watching a mere blank display in an experiment in which the experimenter is not present, unlike in a lab experiment.

The Validation of Stimuli

The validation of our stimulus set was verified in Kawahara et al.'s (2021) study, which investigated cross-cultural audiovisual emotion perception. Overall, there was no remarkable difference between Japanese and Dutch stimuli. For auditory stimuli, the average fundamental frequency (f_0) was higher in Japanese than in Dutch for the happy voice stimuli ($z = -3.36$, $p < 0.001$), but not for the angry voice stimuli (Table 1). Considering that both Japanese and Dutch adult participants in Kawahara et al. (2021) responded to their ingroup voice stimuli more correctly than to their outgroup stimuli, this difference reflected their natural expressions in each culture. For visual stimuli, a certified FACS (Facial Action Coding System; Ekman and Friesen, 1978) coder coded all activated AUs during each stimulus. There was no difference in activated AUs except for AU17⁵ in angry faces ($z = -3.00$, $p = 0.01$) between Japanese and Dutch visual stimuli. Thus, the stimulus set was validated.

⁵According to EMFACS (Friesen and Ekman, 1984), the activation of AU 17 is related to negative expressions such as distress and rage. Considering this fact, it is possible that the activation of AU 17 leads to the judgment of facial expressions as negative. Nevertheless, for stimuli in the present study (Kawahara et al., 2021), we confirmed that both Japanese and Dutch adult participants' accuracy of perceiving angry faces did not differ between stimulus cultures. Therefore, we consider that the frequency of activation in AU 17 did not significantly impact participants' judgment in the present study.

⁴<https://crowdfunder.co.jp/>

TABLE 1 | The average fundamental frequency (f0) of auditory stimuli (Hz).

| | Angry | Happy |
|----------|-------|-------|
| Japanese | 242.8 | 336.9 |
| Dutch | 233.1 | 261.4 |

Apparatus

Web Data

Participants used their own earphones or headphones to listen to auditory stimuli and their own computers to watch visual stimuli and control the browser experiment program. We asked participants to use a computer monitor and earphones (or headphones) and recommended that they use the latest version of Google Chrome. We did not specify the models of the devices.

The resolution of participant displays ranged from 915×515 to 1920×1080 . The participants' used Windows (45 participants), Mac OS (3), Android (2), and iOS (1) operating systems and Google Chrome (27), Microsoft Edge (19), Microsoft Internet Explorer (4), and Safari (1) web browsers.

Lab Data

Researchers provided headphones (SONY MDR-ZX660) (used at a comfortable listening level) to present auditory stimuli and computers (Latitude 3540, Dell) to present visual stimuli and control the experiment program using Hot Soup Processor (Onion Software).

Procedure

Web Data

The flow of the procedure is shown in **Figure 1**. Before the experiment, child participants' parents and adult participants received an instructions and documents file that included how to participate in the event and research brief. At the starting time, each participant and their parent joined the Zoom meeting room. The experimenter and the staff communicated with each participant using their web cameras and microphones to help participants relax. After a short communication, the experimenter provided attendees with instructions (e.g., not to click the web browser back button during the experiment, not to influence their children's responses), checked that participants understood the positions of keys for response (D and K) on their keyboards, and guided them to the experiment web page by providing the URL link in the meeting room chatbox. After checking that each participant succeeded in accessing the experiment page, the experimenter instructed each participant to quit the meeting room to avoid low internet connection speeds during the experiment. They were also instructed to return to the same meeting room if they had any problems or reached the browser experiment's final display.

In the browser experiment session, participants' parents proceeded with the experiment by themselves following instructions on the display. We used Gorilla Experiment Builder to control the experimental program and collect data. The browser experiment session consisted of a preparation section, the auditory task, the visual task, and a questionnaire. In the preparation section, participants' parents gave informed

consent and indicated that environment requirements were met (sufficient device battery, headphones or earphones connection, environmental silence, web browser maximization) using checkboxes. Next, parents checked the sound volume with child participants, MP4 file playback, and keyboard operation following displayed instructions. After preparation, participants engaged in the auditory task and the visual task in each task section. The order of tasks was counterbalanced. At the beginning of each task section, participants watched a task instruction movie that included a simple speaking animation describing the task. The flow of each task is shown in **Figure 1** (see panels 2 and 3, Auditory Task and Visual Task).

In the auditory task, participants were instructed to listen to a voice and judge whether the speaker was angry or happy. A fixation point was displayed at the center of the monitor for 500 ms, after which an auditory stimulus was presented. When the response alternatives written in hiragana characters⁶ were displayed, participants responded by pressing D or K keys (i.e., the allocation of response alternatives was counterbalanced). Five hundred ms after participant's response, the next test trial began, for a total of 32 trials. In the visual task, participants were instructed to observe the face of a (muted) speaker and judge whether they were angry or happy. A fixation point was displayed at the center of the monitor for 500 ms, and each visual stimulus was presented successively. As with the auditory task, responses were indicated by pressing keys. 500 ms blank displays were inserted between trials, for 32 trials. For each task, the main trials were conducted following two practice trials.

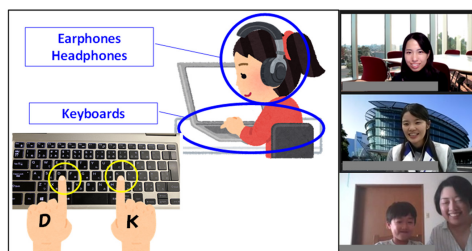
After these tasks, in response to the questionnaire, participants' parents reported problems during the experiment, whether participants had worn earphones, headphones, or had used other devices, whether parents had instructed their children to press a specific key during the main trials ("Did you ask your child to press any specific key during a task, for example, by saying 'Press this key?'"); if they had any concerns, they were asked to fill out a form. According to the questionnaire, we confirmed that all participants included in the analysis had worn earphones or headphones, and that no parent instructed their child to press any specific key. None of parents reported any problems and concerns related with the tasks.

Adult participants similarly joined a Zoom meeting room before the browser experiment and received instructions. They were to proceed with the browser experiment by themselves and return to the same meeting room if they had any problems or had reached the last display in the browser experiment.

The procedure was similar to that of the child participants except that for adult participants the instructions were rewritten

⁶We used response alternatives written in hiragana characters (angry: おこっている[okotteiru]; happy: よろこんでいる[yorokondeiru]) because both young children and adults can read them easily. According to previous studies on Japanese children's literacy, more than half of 3–4-year-olds could read most of hiragana characters (Kakihana et al., 2009), and about 90% of children in 5-year-old classes could read them (Ota et al., 2018) without formal education. Since Japanese hiragana characters have high transparency (each hiragana character corresponds to a syllable in Japanese phonology), Japanese children who know hiragana characters can be assumed to also read words. Considering Japanese children's literacy, we inferred that the participants in this study could read response alternatives.


Video Chat Communication & Instruction



Browser Experiment

① Preparation

- Environment Check
- ☐ Is your device connected to power supply or charged?
 - ☐ Are your earphones (headphones) connected to your device properly?
 - ☐ Are you in a quiet place?
 - ☐ Is your web browser maximized?

- Video and Audio Check
- ① Can you watch the test video?
Can you hear the voice at comfortable volume?
Please check it by a caregiver and a child, respectively.
- 
- ② On wearing earphones (headphones), please adjust the sound volume to hear the voice properly.

② Auditory Task*



③ Visual Task*



*The order of tasks was counterbalanced

④ Questionnaire

Did you have any problems during the experiment?

- ☐ No
- ☐ Yes

Which device did you use?

- ☐ Earphones
- ☐ Headphones
- ☐ Other

FIGURE 1 | Experiment flow (The instructions and alternatives were written in Japanese in the actual experiment).

(e.g., converting some hiragana characters to kanji characters for readability⁷), and they were not asked a question about parents' instruction ("Did you ask your child to press any specific key during a task, for example, by saying 'Press this key?'" in the questionnaire after the task.

Lab Data

The experiment was conducted in an experimental room at the Miraikan for the child participants and in an experimental room at the Tokyo Woman's Christian University for the adult participants. The procedure was almost the same as with the web experiment but with three slight differences. First, in the (Kawahara et al., 2021) lab experiment, auditory and visual tasks were conducted after audiovisual emotional perception tasks in which participants judged speakers' emotions after being presented with face and voice simultaneously. Participants in the web experiment did not engage in audiovisual emotion perception tasks like those in the lab experiment. We cannot rule out the priming effect in the lab data induced by the audiovisual stimuli that had been presented before. However, considering that the number of presentations of "angry" and "happy" stimuli was the same in the audiovisual emotion perception task, a response bias is not possible. Second, cards showing the alternatives ("angry" and "happy" written in hiragana characters) were put on a keyboard in the lab experiment; these alternatives were shown on the display in the web experiment. Third, the experimenter was physically present next to each participant and controlled the experiment program throughout the experiment session in the lab experiment. The presentation of stimuli was controlled using the Hot Soup Processor (Onion Software). These differences were summarized in **Figure 2**.

RESULTS

We calculated the rate of correct responses for each participant. Then, this rate was arcsine transformed to increase the normality of its distribution (accuracy). To investigate whether the experiment method affected different aged participants' performance differently, we conducted a 2 (method: web, lab) \times 3 (age group: younger child, older child, adult) \times 2 (stimulus culture: Japanese, Dutch) mixed-factorial ANOVA on accuracy for each task.

Auditory Task

The results for the auditory task are shown in **Figure 3**. In the auditory task, the main effects of method ($F(1, 257) = 2.07$, $p = 0.151$, $\eta_p^2 = 0.008$), the interaction of method and age group ($F(2, 257) = 0.09$, $p = 0.917$, $\eta_p^2 = 0.001$), of method and stimulus culture ($F(1, 257) = 0.07$, $p = 0.789$, $\eta_p^2 < 0.001$), and of the second-order interaction of method, age group, and stimulus culture ($F(2, 257) = 0.22$, $p = 0.800$, $\eta_p^2 = 0.002$) were not significant. Thus, the results of the auditory task

using online tools were not significantly different from those of the lab experiment.

As for other factors, results showed significant main effects for age group ($F(2, 257) = 31.56$, $p < 0.001$, $\eta_p^2 = 0.20$). The *post hoc* analysis (Shaffer's Modified Sequentially Rejective Bonferroni Procedure) revealed that the older child group of participants responded correctly to more stimulus than the younger child group and that adult participants responded correctly more often than younger and older child participants ($ps < 0.001$). The main effect of stimulus culture was also significant ($F(1, 257) = 281.61$, $p < 0.001$, $\eta_p^2 = 0.52$), showing that participants responded correctly to the Japanese voice more often than to the Dutch voice. Interaction between age group and stimulus culture was marginally significant ($F(2, 257) = 2.89$, $p = 0.057$, $\eta_p^2 = 0.02$). To check whether the impact of stimulus culture was different among age group, we conducted a simple main effect analysis. The simple main effect analysis also showed that all groups selected more correct answers in responses to the Japanese voice than in response to the Dutch voice (Younger child group: $F(1, 119) = 157.56$, $p < 0.001$; Older child group: $F(1, 91) = 97.22$, $p < 0.001$; Adult group: $F(1, 47) = 59.11$, $p < 0.001$). Moreover, a simple main effect of age was significant for both the Japanese voice ($F(2, 257) = 9.01$, $p < 0.001$) and the Dutch voice ($F(2, 257) = 42.04$, $p < 0.001$). *Post hoc* analysis (Shaffer's Modified Sequentially Rejective Bonferroni Procedure) revealed that older child participants responded correctly to the Japanese voice more than the younger child group ($p = 0.049$), and that adult participants responded correctly more often than younger ($p < 0.001$) and older child participants ($p = 0.034$). A similar accuracy difference between age groups was observed with the Dutch voice. Older child participants responded correctly to the Dutch voice more often than the younger child group, and adult participants responded correctly more often than younger and older child participants ($ps < 0.001$). All age groups responded correctly more often to the Japanese voice than the Dutch voice, and accuracy with both voices increased with age.

To further examine the marginal interaction between age group and stimulus culture, we conducted a two-way ANOVA (method \times age group) on the ingroup advantage. This was calculated by subtracting the accuracy of Dutch voices from that of Japanese voices. The main effect of age groups was marginally significant ($F(2, 257) = 2.89$, $p = 0.057$, $\eta_p^2 = 0.02$). *Post hoc* analysis (Shaffer's Modified Sequentially Rejective Bonferroni Procedure) showed that the difference between younger children and adults was marginally significant ($p = 0.051$), suggesting that the ingroup advantage in younger children was more salient than that in adults. The differences between other pairs were not significant (younger child group – older child group: $p = 0.149$, older child group – adult group: $p = 0.393$). The main effects of method ($F(1, 257) = 0.07$, $p = 0.789$, $\eta_p^2 < 0.001$) and the interaction of method and age group ($F(2, 257) = 0.22$, $p = 0.800$, $\eta_p^2 = 0.002$) were not significant.

To further examine the effect of method on accuracy, we conducted a Bayesian repeated measures ANOVA on accuracy using JASP (2018) with default prior scales. **Table 2** shows the inclusion probabilities and the inclusion Bayes factor (Clyde et al., 2011; van den Bergh et al., 2019). The inclusion Bayes

⁷We did it because it is easier for native Japanese speaking adults to read sentences such as instructions that use both kanji and hiragana than sentences written in hiragana characters only.


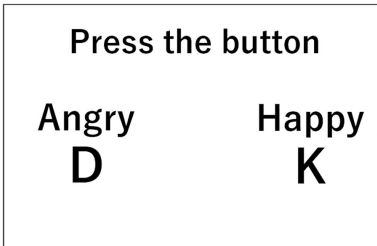
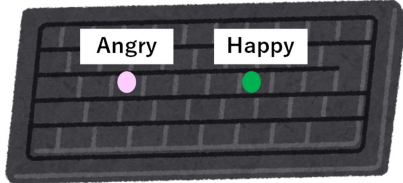
| | Web Data | Lab Data |
|---|---|---|
| 1 Components of tasks | Communication with an experimenter ↓ Instruction Auditory Task ↓ Instruction Visual Task | Communication with an experimenter ↓ Instruction Audiovisual Task ↓ Instruction Auditory Task ↓ Instruction Visual Task |
| 2 Presentation of response alternatives* | On a display (Check through a video chat ) On each trial  | On a keyboard  |
| 3 Experimenter | -was absent during tasks (parents were present) | -was present during tasks |

FIGURE 2 | The differences in procedures between web and lab data.

factors reflect the average across possible models and reveal whether models with a particular predictor are more likely to have produced the observed data than those without. This approach is especially useful when the number of potential variables under consideration is large. The Bayesian ANOVA revealed that the $BFinclusion$ values of the effect of method ($BFinclusion = 0.165$), the interaction effect between method and stimulus culture

($BFinclusion = 0.111$), the interaction effect between method and age group ($BFinclusion = 0.069$), and the second-order interaction of method, age group, stimulus culture, and age ($BFinclusion = 0.004$) were all small, supporting no effect of the difference between the web and lab experiments. Additionally, the data provide strong evidence for the effects of age group and stimulus culture ($BFinclusions > 100$), although they are not

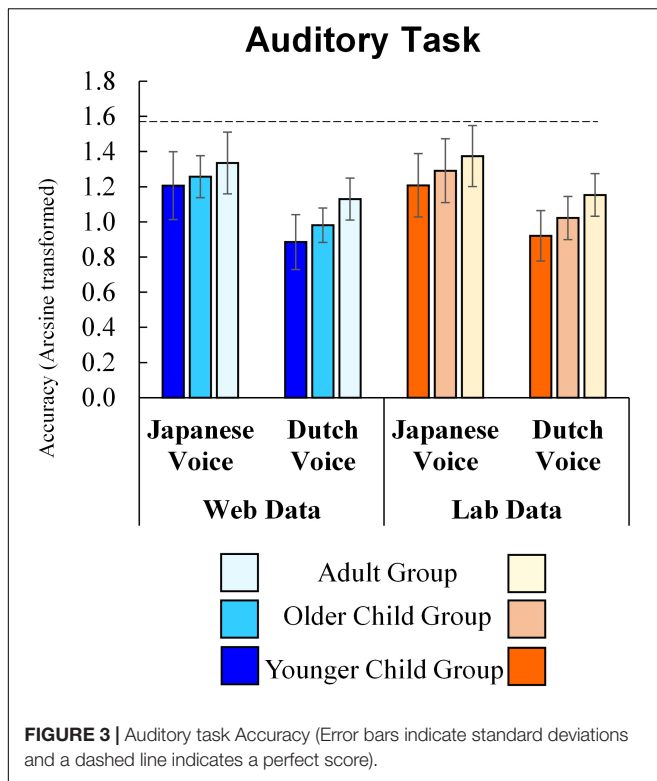


TABLE 2 | Evidence for the presence of particular effects in the accuracy of the auditory task (Data averaged over all the models including/excluding a particular predictor).

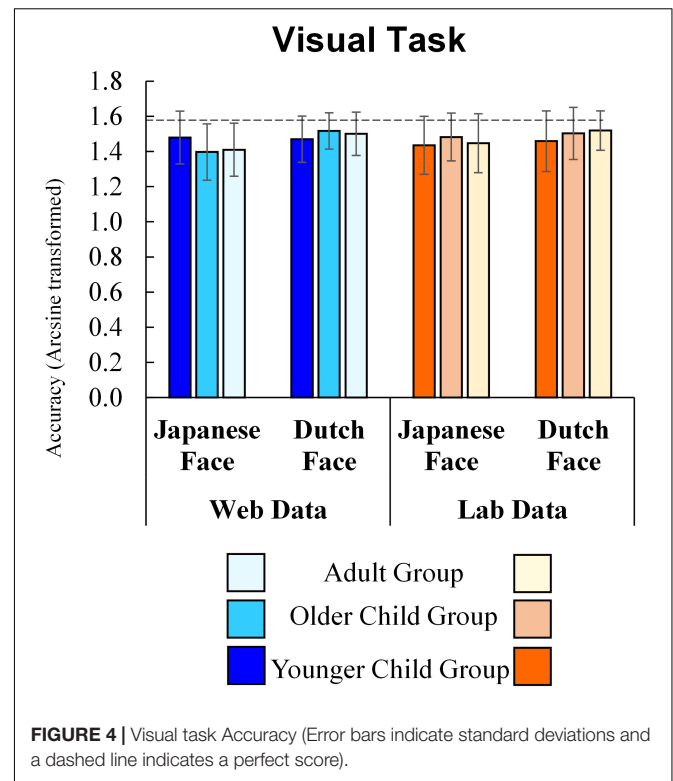
| | P(incl) | P(excl) | P(incl data) | P(excl data) | BFincl |
|--|---------|---------|---------------|---------------|----------|
| Stimulus Culture | 0.737 | 0.263 | 1.000 | 0.000 | ∞ |
| Age Group | 0.737 | 0.263 | 1.000 | 0.000 | ∞ |
| Method | 0.737 | 0.263 | 0.316 | 0.684 | 0.165 |
| Stimuli \times Age Group | 0.316 | 0.684 | 0.336 | 0.664 | 1.095 |
| Stimuli \times Method | 0.316 | 0.684 | 0.049 | 0.951 | 0.111 |
| Method \times Age Group | 0.316 | 0.684 | 0.031 | 0.969 | 0.069 |
| Stimuli \times Method \times Age Group | 0.053 | 0.947 | > 0.001 | 1.000 | 0.004 |

sufficiently informative to allow a strong conclusion about the effect of the interaction between age group and stimulus culture (BFinclusion = 1.095).

Visual Task

The results of the visual task are shown in **Figure 4**. In the visual task, the main effects of method ($F(1, 257) = 0.36, p = 0.551, \eta_p^2 = 0.001$), the interaction of method and age group ($F(2, 257) = 1.10, p = 0.335, \eta_p^2 = 0.008$), method and stimulus culture ($F(1, 257) = 1.00, p = 0.320, \eta_p^2 = 0.004$), and the second-order interaction of method, age group and stimuli ($F(2, 257) = 1.99, p = 0.138, \eta_p^2 = 0.015$) were not significant. Thus, the results of the visual task using online tools were not significantly different from those of the lab experiment.

Results indicated a significant main effect for stimulus culture ($F(1, 257) = 13.92, p < 0.001, \eta_p^2 = 0.051$), showing that



participants responded correctly to the Dutch face more often than to the Japanese face. Interaction between age group and stimulus culture ($F(2, 257) = 3.05, p = 0.049, \eta_p^2 = 0.023$) was also significant, but the main effect for age group was not ($F(2, 257) = 0.19, p = 0.829, \eta_p^2 = 0.001$). A simple main effect analysis revealed that older children ($p = 0.007$) and adults ($p = 0.003$) responded correctly to the Dutch face more often than to the Japanese face, while younger children's accuracy did not differ between stimulus cultures ($p = 0.744$). The accuracy for faces did not differ among age groups both for Japanese ($p = 0.624$) and Dutch stimuli ($p = 0.165$).

To further examine the effect of method on accuracy, similar to the auditory task, we conducted a Bayesian repeated measures ANOVA using JASP with default prior scales. **Table 3** shows the inclusion probabilities and inclusion Bayes factor. The Bayesian ANOVA revealed that the effect of method type (BFinclusion = 0.084), the interaction effect between method and stimulus culture (BFinclusion = 0.083), the interaction effect between method and age group (BFinclusion = 0.033), and the second-order interaction of method, age group, stimulus culture, and age (BFinclusion = 0.006) were all small, supporting no effect of the difference between the web and lab experiments. Consistent with the results of classical ANOVA, the data provided moderate evidence for the effect of stimulus culture (BFinclusion = 5.893). However, the effect of the interaction between age group and stimulus culture was not informative (BFinclusion = 0.128).

Thus, results showed that the experiment method (web or lab) did not affect participants' performance in either the auditory task

TABLE 3 | Evidence for the presence of particular effects in the accuracy of the visual task (Data averaged over all the models including/excluding a particular predictor).

| | P(inkl) | P(excl) | P(inkl data) | P(excl data) | BFincl |
|------------------------------|---------|---------|---------------|---------------|--------|
| Stimulus Culture | 0.737 | 0.263 | 0.943 | 0.057 | 5.893 |
| Age Group | 0.737 | 0.263 | 0.300 | 0.700 | 0.153 |
| Method | 0.737 | 0.263 | 0.191 | 0.809 | 0.084 |
| Stimuli × Age Group | 0.316 | 0.684 | 0.056 | 0.944 | 0.128 |
| Stimuli × Method | 0.316 | 0.684 | 0.037 | 0.963 | 0.083 |
| Method × Age Group | 0.316 | 0.684 | 0.015 | 0.985 | 0.033 |
| Stimuli × Method × Age Group | 0.053 | 0.947 | >0.001 | 1.000 | 0.006 |

or the visual task; that is, the web data obtained in the present study did not differ from our previously obtained lab data. That is, our method can enable researchers to obtain data that would be comparable to those of a laboratory experiment in the emotion perception tasks.

Reaction Time

Here we reported reaction times in the online experiment. Since we did not measure reaction times in the lab experiment, we cannot compare them between methods. Moreover, we did not instruct participants to respond to each stimulus as quickly as they could. Thus, the reaction times data here are for reference only. Nevertheless, this is useful as data of online developmental experiment, and it also enable us to investigate whether we could find age difference in performance for the visual task, in which no age difference in accuracy was found due to the ceiling effect.

We showed average reaction times of each task in **Figure 5**. We excluded outlier reaction time data (each participant's average reaction time ± 2.5 SD), while including trials in which participants pressed the wrong key, considering the following reasons. First, children's responses classified as "incorrect responses" may include the results of their careful consideration. Second, given the age differences in accuracy, the number of correct responses, that is, the number of trials included in the analysis differed among age groups in the auditory task. We conducted a 3 (age group) \times 2 (stimulus culture) mixed-factorial ANOVA on reaction time for each task. In the auditory task, the main effects for age group ($F(2, 48) = 4.91, p = 0.011, \eta_p^2 = 0.17$) was significant. The *post hoc* analysis (Shaffer's Modified Sequentially Rejective Bonferroni Procedure) revealed that adult group of participants responded more quickly than the younger ($p = 0.019$) and older child groups ($p = 0.019$). There was no significant difference between the younger and older child groups ($p = 0.587$). The main effect of stimulus culture was also significant ($F(1, 48) = 27.67, p < 0.001, \eta_p^2 = 0.37$), showing that participants responded more quickly to the Japanese voice than to the Dutch voice. Interaction between age group and stimulus culture was also significant ($F(2, 48) = 3.91, p = 0.027, \eta_p^2 = 0.14$). To check whether the impact of stimulus culture was different among age groups, we conducted a simple main effect analysis. The simple main effect analysis also showed that all groups yielded faster responses to the Japanese voice than in response to

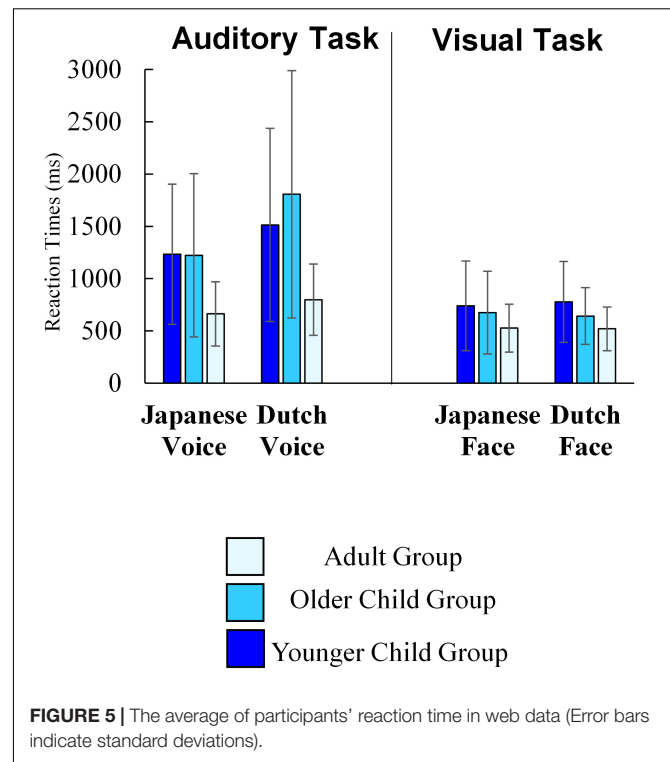


FIGURE 5 | The average of participants' reaction time in web data (Error bars indicate standard deviations).

the Dutch voice (Younger child group: $F(1, 20) = 6.01, p = 0.024$; Older child group: $F(1, 13) = 17.37, p = 0.011$; Adult group: $F(1, 15) = 8.51, p = 0.011$). Moreover, a simple main effect of age was significant for both the Japanese voice ($F(2, 48) = 4.35, p = 0.018$) and the Dutch voice ($F(2, 48) = 5.08, p < 0.001$). *Post hoc* analysis (Shaffer's Modified Sequentially Rejective Bonferroni Procedure) revealed that adult participants responded to Japanese voice faster than older children ($p = 0.029$) and younger children ($p = 0.029$), and adult participants responded to Dutch voice faster than older children ($p = 0.011$) and younger children ($p = 0.022$). There was no significant difference between the younger and older child groups for both Japanese voice ($p = 0.962$) and Dutch voice ($p = 0.351$). We conducted a one-way ANOVA on the ingroup advantage to further examine the interaction between age group and stimulus culture. This was calculated by subtracting the reaction time to Japanese voices from that of Dutch voices. The main effect of age groups was significant ($F(2, 48) = 3.91, p = 0.027, \eta_p^2 = 0.01$). *Post hoc* analysis (Shaffer's Modified Sequentially Rejective Bonferroni Procedure) showed that the ingroup advantage in older children was larger than that in both adults ($p = 0.025$) and younger children (marginally significant; $p = 0.053$). The difference between younger children and adults was not significant ($p = 0.338$). Thus, the ingroup advantage based on reaction times was most salient in older children, unlike the analysis of accuracy.

In the visual task, the main effects of age group ($F(2, 48) = 2.18, p = 0.124, \eta_p^2 = 0.08$) and of stimulus culture ($F(1, 48) < 0.01, p = 0.991, \eta_p^2 < 0.001$), and the interaction between them ($F(2, 48) = 0.84, p = 0.438, \eta_p^2 = 0.03$) were not significant. Thus, all groups responded visual stimulus quickly and there was

no difference among age groups and between stimulus culture. This is consistent with that high accuracy was observed in all age groups. Taken together, there was no differences among age groups both in accuracy and reaction times in the visual task.

DISCUSSION

The purpose of this study was to investigate the validity of online developmental studies through an emotion perception experiment. To that end, we conducted an experiment controlled by Gorilla Experiment Builder with child and adult participants who engaged in simple auditory and visual emotion perception tasks following communication and instruction through Zoom. As predicted, results showed no significant differences in participants' performance between our web data results and out lab data results (Kawahara et al., 2021). In the auditory task, we found performance differences between age groups (between-subject factor) and better performance with stimuli spoken in their native language (within-subject factor). These findings were consistent with previous laboratory studies reporting performance improvements with age (Sauter et al., 2013; Chronaki et al., 2014) and superior perception of vocal emotional expression with native language stimulus (Sauter et al., 2010). In the visual task, accuracy was high and near-perfect among all age groups both in our web and lab data. To date, although online experiments have replicated laboratory experiment results with adult participants (e.g., Crump et al., 2013), developmental studies with child participants have been limited. By including child participants in our online study of emotional perception tasks, the present study adds new evidence regarding the validity of data collected in online developmental studies.

Notably, accuracy of perception of vocal expression in the lab experiment was replicated in the web experiment even though participants were not required to use a specific device and were allowed to use any earphones or headphones. The results of the present study may relieve researchers' hesitation to conduct developmental experiments online, at least in the field of auditory emotional perception. Of course, we need additional examination to determine the suitability of online platforms for other types of auditory perception research. In tasks such as phoneme perception, judgment of speaker identities, or perceiving vocal expression from among multiple choices, participants' responses may be affected by the devices they use (see Woods et al., 2017).

We cannot strongly conclude that online developmental research is valid for the task of perceiving facial expressions because we observed a ceiling effect; that is, performance was near perfect among all age groups in the present task. Our stimuli for the visual task were quite clear—Tanaka et al. (2010) originally created them by adding random dynamic noises to be degraded—and only two response alternatives (angry or happy) were available. The reasoning behind Kawahara et al.'s (2021) use of visual stimuli without noises was to avoid unpleasantness for the children, and so the present study followed that procedure. However, to investigate the impact of browser experiments on the presentation of visual stimuli in detail, we should conduct online

studies using low intensity facial expressions, with more variety of emotions, or with smaller sized pictures in the future.

Although our main purpose was to investigate the validity of data obtained through online experiments, our data provide interesting findings on perceptual development. First, in the auditory task, the difference in the ingroup advantage was marginally significantly different between younger children and adults. That is, the ingroup advantage in younger children may be more salient than that in adults. This tendency may be related to the findings that young children prefer people who spoke their native language (e.g., Kinzler et al., 2007). Second, in the visual task, participants gave more correct responses to Dutch faces than to Japanese faces. These results are unexpected considering that previous studies have insisted on the ingroup effect in facial recognition tasks (e.g., Elfenbein and Ambady, 2002). However, more recent studies have demonstrated that Japanese raters did not show the ingroup advantage in the perception of facial expressions (Matsumoto et al., 2009; Hutchison et al., 2018). Overall, Japanese people's facial expressions may not necessarily be perceived correctly by ingroup members. Our results of the visual task may also reflect this. Moreover, the interaction between age group and stimulus culture on accuracy was also significant in the visual task, suggesting that older children and adults responded more correctly to Dutch facial expressions than to Japanese facial expressions. These may be interesting if Japanese people judge their outgroup facial expressions more accurately compared to Japanese stimuli with age. Japanese children may come to know that Japanese people tend to conceal their true emotions (e.g., Matsumoto, 1990) and that they inhibit their facial expressions, and this may cause "outgroup advantage." This may lead to Japanese people's tendency to prioritize voice in audiovisual emotion perception as for Japanese stimuli shown in previous studies (Yamamoto et al., 2020; Kawahara et al., 2021). We cannot clarify this speculation based on the present study because the Bayesian ANOVA did not provide strong evidence for the interaction between age group and stimulus culture in both tasks. However, we need to investigate these perceptual developmental suggestions in the further study.

We showed new possibilities for using simple, general (not specialized for children) online tools that may enable researchers to move their laboratory studies online. However, we should clarify the limitations of the methods presented here. First, this study targeted children who could be instructed verbally and could respond by pressing keys on a keyboard. Considering that Japanese preschool children read written words from relatively early on, we did not check the level of literacy for each participant. However, checking this would be important for researchers to apply this method to children living in various environments. For studies targeting preliterate children as participants, researchers should select a video-recorder type experiment model and record participants' oral or pointing responses. Second, we had to rely on parents' self-reports and could not independently check children's actual states during the browser experiment because we instructed parents to turn off their web cameras. Although we ensured that parents did not ask their children to respond in line with parents' answers during the main trials by questionnaires,

we did not have a way to confirm this was the case. Moreover, it is possible that parents would have given their children instructions without being aware of it. This could be avoided by keeping the web cameras on during the experiment. However, this could affect the quality of the presentation of the stimuli due to internet connection speed issues. These are the trade-offs, and in the present study we gave prioritized the quality of the stimuli. Such choices should be made in accordance with the aim of each study. Moreover, we should not forget the burden on parents during online experiments, and minimizing this burden should be considered when designing online experiments. In addition, as described in the section “Materials and Methods,” our web experiment procedure was not exactly the same as that of our laboratory experiment (**Figure 2**). Nevertheless, there were no significant differences between our web data and lab data in the present study, which suggests that the difference in procedure does not have a critical impact on the results in the experiments investigating the development of emotion perception.

As an online experiment research, the procedure of the present study has two particularities. First, child participants’ parents take on the “experimenter” role. Second, researchers and participants communicate with each other through video chat before the tasks. Previous online psychological experiment studies with adults have pointed out high dropout rates (e.g., Reips, 2002), large variances due to various environments among participants, and difficulties in a between-subject design study. On the contrary, it is worth noting that the methods we adopted resulted in very few cases of participant data being excluded. Moreover, the variance of performance seemed to be similar to that of the lab experiment even though participants used their laptops and earphones (headphones). Our only requests before the experiment were to enter the video chat at the appointed time and to prepare earphones or headphones. Considering the effort involved in making an appointment with each participant, and in instructing both parents and child participants to proceed with the experiments, our method does not have benefits such as large data collecting in a short period, unlike usual crowdsourced online experiments. Nevertheless, the results of the present study suggest that this effort can reduce issues associated with online research, such as a dropout rate and variance of data. Given that even adult participants engaged in tasks seriously without an experimenter, video chat communication before the main experiment may be specifically effective. Even though our method does not have the aforementioned benefits associated with crowdsourced online research, we regard its biggest advantage to be the fact that both experimenters and participants are not affected by geographical constraints. In fact, as long as they have an internet connection, researchers can conduct studies with people living in various countries and continue to collect data even under a pandemic.

We should note that the reproducibility of results in online experiments may depend on indices. We used the rate of correct responses, and we did not compare other indices such as reaction times or fixations. Given that our web data showed that all age groups responded more quickly to the Japanese stimuli in the auditory task, reaction time may be used even in online experiments. However, unlike the results for accuracy, we did

not find age differences between child groups for reaction times. Moreover, while we observed a salient ingroup advantage in younger children compared with adults (although the tendency was marginally significant) for accuracy, this was not reflected in reaction times. Rather, for reaction times, we found that older children’s ingroup advantage was more salient than the other two age groups. Since we do not have reaction time data of our lab experiment, it remains unclear whether such tendency is observed also in a lab experiment or is unique to an online experiment. As a limitation of online research, one previous study pointed out the difficulty in controlling a short presentation of stimuli such as a masked priming procedure (Crump et al., 2013). Another study investigating the contrast threshold reported a high rate of data exclusions due to each participant’s experimental environment (Sasaki and Yamada, 2019). Further studies are needed to examine which indices and tasks are adequate for online experiments. Despite the limitations, we demonstrated that online experiments are useful for child research using auditory and visual (movie) stimuli. Combinations of online tools will lead researchers to new developmental research styles. Moreover, due to the validity of this online research using unimodal auditory and visual stimuli, application to future audiovisual perception research is expected.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Tokyo Woman’s Christian University Research Ethics Committee. Written informed consent or digital informed consent to participate in this study was provided by the participants’ legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

HY was involved in designing the web experiment programs, collecting the data in the web experiment, analyzing the data, and drafting the manuscript. MK was involved in collecting the data and designing programs in the laboratory experiment. AT was involved in the creation of stimuli. All authors were involved in the experimental design, interpretation of the results, revised, and approved the final version of the manuscript.

FUNDING

This work was supported by JSPS KAKENHI (No. 20J01281) and Grant-in-Aid for Scientific Research on Innovative Areas No. 17H06345 “Construction of the Face-Body Studies in Transcultural Conditions”.

ACKNOWLEDGMENTS

We thank all the participants and staff members of the National Museum of Emerging Science and Innovation (Miraikan).

REFERENCES

- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: an online behavioral experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x
- Carbon, C.-C. (2020). Wearing face masks strongly confuses counterparts in reading emotions. *Front. Psychol.* 11:566886. doi: 10.3389/fpsyg.2020.566886
- Chronaki, G., Hadwin, J. A., Garner, M., Maurage, P., and Sonuga-Barke, E. J. S. (2014). The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. *Br. J. Dev. Psychol.* 33, 218–236. doi: 10.1111/bjdp.12075
- Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *J. Comput. Graph. Stat.* 20, 80–101. doi: 10.1198/jcgs.2010.09049
- Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS One* 8:e57410. doi: 10.1371/journal.pone.0057410
- de Leeuw, J. R., and Motz, B. A. (2016). Psychophysics in a web browser? Comparing response times collected with javascript and psychophysics toolbox in a visual search task. *Behav. Res. Methods* 48, 1–12. doi: 10.3758/s13428-015-0567-2
- Ekman, P., and Friesen, W. V. (1978). *Facial Action Coding System: Investigator's Guide*. Palo Alto, CA: Consulting Psychologists Press.
- Elfenbein, H. A., and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychol. Bull.* 128, 203–235. doi: 10.1037/0033-2909.128.2.203
- Friesen, W. V., and Ekman, P. (1984). *EMFACS-7. Unpublished manuscript*. Human Interaction Laboratory. San Francisco, CA: University of California.
- Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. (2012). Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychon. Bull. Rev.* 19, 847–857. doi: 10.3758/s13423-012-0296-9
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). *The Weirdest People in the World? (RatSWD Working Paper Series, 139)*. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD).
- Hutchison, A., Gerstein, L., and Kasai, M. (2018). A cross-cultural comparison of u.s. and japanese trainees' emotion-recognition ability. *Jpn. Psychol. Res.* 60, 63–76. doi: 10.1111/jpr.12182
- JASP (2018). *JASP (Version 0.9) [Computer software]*. Available online at: <https://jasp-stats.org> (accessed June 21, 2021).
- Kakihana, S., Ando, J., Koyama, M., Iitaka, S., and Sugawara, I. (2009). Cognitive factors relating to the development of early literacy in the Kana syllabary. *Jpn. J. Educ. Psychol.* 57, 295–308. doi: 10.5926/jjep.57.295
- Kawahara, M., Sauter, D., and Tanaka, A. (2021). Culture shapes emotion perception from faces and voices: changes over development. *Cogn. Emot.* doi: 10.1080/02699931.2021.1922361 [Epub ahead of print].
- Kinzler, K. D., Dupoux, E., and Spelke, E. S. (2007). The native language of social cognition. *Proc. Natl. Acad. Sci. U.S.A.* 104, 12577–12580. doi: 10.1073/pnas.0705345104
- Klindt, D., Devaine, M., and Daunizeau, J. (2017). Does the way we read others' mind change over the lifespan? Insights from a massive web poll of cognitive skills from childhood to late adulthood. *Cortex* 86, 205–215. doi: 10.1016/j.cortex.2016.09.009
- Laeng, B., Kiambarua, K. G., Hagen, T., Bochynska, A., Lubell, J., Suzuki, H., et al. (2018). The “face race lightness illusion”: an effect of the eyes and pupils? *PLoS One* 13:e0201603. doi: 10.1371/journal.pone.0201603
- Lavan, N., Burston, L. F. K., and Garrido, L. (2018). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *Br. J. Dev. Psychol.* 110, 576–593. doi: 10.1111/bjop.12348
- Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motiv. Emot.* 14, 195–214. doi: 10.1007/BF00955569
- Matsumoto, D., Olide, A., and Willingham, B. (2009). Is there an ingroup advantage in recognizing spontaneously expressed emotions? *J. Nonverbal. Behav.* 33:181. doi: 10.1007/s10919-009-0068-z
- McPherson, M. J., and McDermott, J. H. (2018). Diversity in pitch perception revealed by task dependence. *Nat. Hum. Behav.* 2, 52–66. doi: 10.1038/s41562-017-0261-8
- Mills, C., and D'Mello, S. (2014). On the validity of the autobiographical emotional memory task for emotion induction. *PLoS One* 9:e95837. doi: 10.1371/journal.pone.0095837
- Miura, A., and Kobayashi, T. (2016). Survey satisficing inflates stereotypical responses in online experiment: the case of immigration study. *Front. Psychol.* 7:1563. doi: 10.3389/fpsyg.2016.01563
- Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., and Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra. Psychol.* 6:17213. doi: 10.1525/collabra.17213
- Ota, S., Uno, A., and Inomata, T. (2018). Attainment level of hiragana reading / spelling in kindergarten children. *Jpn. J. Logop. Phoniatr.* 59, 9–15. doi: 10.5112/jjlp.59.9
- Paolacci, G., and Chandler, J. (2014). Inside the turk: understanding mechanical turk as a participant pool. *Psychol. Sci.* 23, 184–188. doi: 10.1177/0963721414531598
- Reips, U. D. (2002). Standards for internet-based experimenting. *Exp. Psychol.* 49, 243–256. doi: 10.1027/1618-3169.49.4.243
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Sasaki, K., and Yamada, Y. (2019). Crowdsourcing visual perception experiments: a case of contrast threshold. *PeerJ* 7:e8339. doi: 10.7717/peerj.8339
- Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2408–2412. doi: 10.1073/pnas.0908239106
- Sauter, D. A., Panattoni, C., and Happé, F. (2013). Children's recognition of emotions from vocal cues. *Br. J. Dev. Psychol.* 31, 97–113.
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (Part 2): assessing the viability of online development research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/opmi_a_00001
- Semmelmann, K., and Weigelt, S. (2017). Online psychophysics: reaction time effects in cognitive experiments. *Behav. Res. Methods* 49, 1241–1260. doi: 10.3758/s13428-016-0783-4
- Sheskin, M., and Keil, F. (2018). TheChildLab.com: a video chat platform for developmental research. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/rn7w5
- Shin, H., and Ma, W. J. (2016). Crowdsourced single-trial probes of visual working memory for irrelevant features. *J. Vis.* 16, 1–8. doi: 10.1167/16.5.10
- Simcox, T., and Fiez, J. A. (2014). Collecting response times using amazon mechanical turk and adobe flash. *Behav. Res. Methods* 46, 95–111. doi: 10.3758/s13428-013-0345-y
- Stewart, N., Chandler, J., and Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends Cogn. Sci.* 21, 736–748. doi: 10.1016/j.tics.2017.06.007
- Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., and de Gelder, B. (2010). I feel your voice. Cultural differences in the multisensory perception of emotion. *Psychol. Sci.* 21, 1259–1262. doi: 10.1177/0956797610380698

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.702106/full#supplementary-material>

- Tran, M., Cabral, L., Patel, R., and Cusack, R. (2017). Online recruitment and testing of infants with Mechanical Turk. *J. Exp. Child Psychol.* 156, 168–178. doi: 10.1016/j.jecp.2016.12.003
- van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E., Derks, K., et al. (2019). A tutorial on conducting and interpreting a bayesian anova in JASP. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/spreb.
- Woods, K. J. P., Siegel, M. H., Traer, J., and McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Atten. Percept. Psychophys.* 79, 2064–2072. doi: 10.3758/s13414-017-1361-2
- Yamamoto, H. W., Kawahara, M., and Tanaka, A. (2020). Audiovisual emotion perception develops differently from audiovisual phoneme perception during childhood. *PLoS One* 15:e0234553. doi: 10.1371/journal.pone.0234553
- Zhou, H., and Fishbach, A. (2016). The pitfall of experimenting on the web: how unattended selective attrition leads to surprising (yet false) research conclusions. *J. Pers. Soc. Psychol.* 111, 493–504. doi: 10.1037/pspa0000056

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Yamamoto, Kawahara and Tanaka. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The Community-Engaged Lab: A Case-Study Introduction for Developmental Science

Judy Liu^{1†}, Scott Partington^{1†}, Yeonju Suh^{1†}, Zoe Finiasz¹, Teresa Flanagan¹, Deanna Kocher¹, Richard Kiely², Michelle Kortenaar³ and Tamar Kushnir^{1*}

¹ Department of Human Development, Cornell University, Ithaca, NY, United States, ² Office of Engagement Initiatives, Cornell University, Ithaca, NY, United States, ³ Sciencenter, Ithaca, NY, United States

OPEN ACCESS

Edited by:

Natasha Kirkham,
Birkbeck, University of London,
United Kingdom

Reviewed by:

Sara Rodriguez-Cuadrado,
Autonomous University of
Madrid, Spain
Joni Tzuchen Tang,
National Taiwan University of Science
and Technology, Taiwan

*Correspondence:

Tamar Kushnir
tk397@cornell.edu

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 27 May 2021

Accepted: 27 July 2021

Published: 19 August 2021

Citation:

Liu J, Partington S, Suh Y, Finiasz Z,
Flanagan T, Kocher D, Kiely R,
Kortenaar M and Kushnir T (2021) The
Community-Engaged Lab: A
Case-Study Introduction for
Developmental Science.
Front. Psychol. 12:715914.
doi: 10.3389/fpsyg.2021.715914

Due to the closing of campuses, museums, and other public spaces during the pandemic, the typical avenues for recruitment, partnership, and dissemination are now unavailable to developmental labs. In this paper, we show how a shift in perspective has impacted our lab's ability to successfully transition to virtual work during the COVID-19 shut-down. This begins by recognizing that any lab that relies on local communities to engage in human research is *itself a community organization*. From this, we introduce a *community-engaged lab* model, and explain how it works using our own activities during the pandemic as an example. To begin, we introduce the vocabulary of mission-driven community organizations and show how we applied the key ideas of mission, vision, and culture to discussions of our own lab's identity. We contrast the community-engaged lab model with a traditional bi-directional model of recruitment *from* and dissemination *to* communities and describe how the community-engaged model can be used to reframe these and other ordinary lab activities. Our activities during the pandemic serve as a case study: we formed new community partnerships, engaged with child "citizen-scientists" in online research, and opened new avenues of virtual programming. One year later, we see modest but quantifiable impact of this approach: a return to pre-pandemic diversity in our samples, new engagement opportunities for trainees, and new sustainable partnerships. We end by discussing the promise and limitations of the community-engaged lab model for the future of developmental research.

Keywords: COVID-19, community engagement, online developmental science, citizen science, research-community partnerships, broader impacts

INTRODUCTION

Developmental science does not happen in a vacuum. Our science crucially depends on our relationship with local communities—and in particular the organizations and spaces where children and families live, work, and play. This includes schools, museums, daycare centers, churches, playgrounds, local businesses, and the many community non-profits that serve children's and families' interests. It is standard practice in our discipline to engage with local organizations when we recruit families to participate in research. It is also standard to include them as part of our plans to disseminate research beyond academic publications. Our impact and success depend on cultivating collaborative research partnerships with our local communities.

We believe research labs can benefit from more explicitly acknowledging this fact. In this paper, we use our own lab as a case study, and argue that a simple shift in perspective can have a positive impact on research, dissemination, and bridge-building between labs and local communities. Our own shift to this perspective began prior to the 2020 coronavirus pandemic, and we believe it allowed us to transition to pandemic-era work with relative ease.

We begin by describing the principles of mission-driven community organizations and how they can be used to create a new model for developmental science labs. We then describe the model of the *community-engaged lab* and contrast it with the standard bi-directional model. Using examples, we demonstrate how this model enabled us to pivot to new recruitment methods, programming, and dissemination in a completely virtual pandemic-era lab. We then present evidence suggesting that community engagement *works*: internal records show that our virtual engagement efforts helped to maintain a representative participant pool on par with our recruitment pre-pandemic, that we have increased opportunities for early-career researchers to get involved in public-facing programs and outreach, and that we have expanded to include engagement with new and different types of community partners. We conclude with discussion of the unique benefits that engagement can provide to our communities and our science, as well as some ways that in-person and digital avenues of community engagement can complement each other in the future.

PRINCIPLES OF MISSION-DRIVEN COMMUNITY ORGANIZATIONS

Community organizations that work with children and families define their purpose and contributions with a *mission*, *vision*, and *values* (Crutchfield and Grant, 2007). The *mission* is an explanation of how the organization's vision will be accomplished, while the *vision* is a statement that describes long-term goals or purpose. A strong mission statement is rooted in the present and is also purpose-driven: it is future-oriented and often a means of achieving a greater vision. A mission statement also typically includes a target audience, the organization's contribution, and factors that distinguish it from other organizations. *Values* are fundamental guiding principles and beliefs that help define an organization's identity and an organization's culture. An effective value statement explicitly states how members of an organization are expected to act toward fellow internal members as well as how the organization will treat the community as a whole.

Successful organizations make their mission, vision, and values explicit (Crutchfield and Grant, 2007). Though we had begun some of these discussions prior to the pandemic, this topic took priority in our discussions when the shut-down occurred. How does a lab like ours develop a mission statement? Our lab conducts basic research in cognitive and social development. It is a place where future scientists are trained, at the undergraduate and graduate level, by participating in the day-to-day work of conducting research, by studying developmental theory, and

by collaborating and exchanging ideas. Thus, it is perhaps obvious that our mission centers around research, teaching, and mentorship. Essential target audiences therefore include students that receive training and mentorship in the lab and the scientific community that we reach through scholarly publications.

Does our mission extend beyond the research, teaching, and mentorship goals of our scientific enterprise? It does if our vision does. Our research program centers around early childhood learning, cognitive, and social-cognitive development from a constructivist perspective. Moreover, for many years, we have been in partnerships with educational institutions—including our local science museum, local schools, and youth programs. By combining these, our explicit statement of vision became: “to empower communities with a holistic understanding of child development, so that every child can actively explore and learn about the world around them, supported by caring adult guidance and the surrounding culture.” This statement reflects what we believe to be fundamental principles of early learning and development, and also reflects what our community partnerships have taught *us* about *their* missions, and our respective contributions as partners.

As is true of other community organizations, our mission and vision are carried out through daily actions, guided by shared values. Again through discussion, we worked to make our values explicit. What we settled on was a culture of *trust*, *collaboration*, and *acceptance* that defines how we interact within our lab and guides ethical action toward our participants, our partners, and others. In our view, trust is the foundation of responsible research conduct, protecting data integrity, and working in teams. Similarly, we view collaboration is the basis of creative scholarship, and involves a willingness to combine strengths, to teach and mentor, and to listen openly in a free exchange of ideas. Finally, acceptance allows expression of different perspectives, intellectual risk-taking, and appreciation of each other as we work toward common goals.

This internal culture informs our relationships with the local community. We trust each other to represent the lab honorably when interacting with participants and their families. Communities trust us to ask how we can best meet their needs and value their assets, and not assume that we know what those needs and assets are (Kretzmann and McKnight, 1993). Trainees in the lab *double as community ambassadors* as they engage in active volunteerism in community organizations and science communication. Several examples are found in sections below. Importantly, in our minds, this culture opens up space for new collaborations, for new ideas for programming, for grants, and for research.

We want to make clear that, at least for us, the importance of making these mission, vision, and value statements explicit was more about the process than the outcome (Ash and Clayton, 2009). Last March, the motivation to think in this way was made urgent by the complete fragmentation of everything that made us a lab prior to the pandemic. Individual students and researchers went home. Children were home. Parents were struggling to work, care for, and educate their children alone. We as a group of researchers were looking for a way to connect, to maintain our

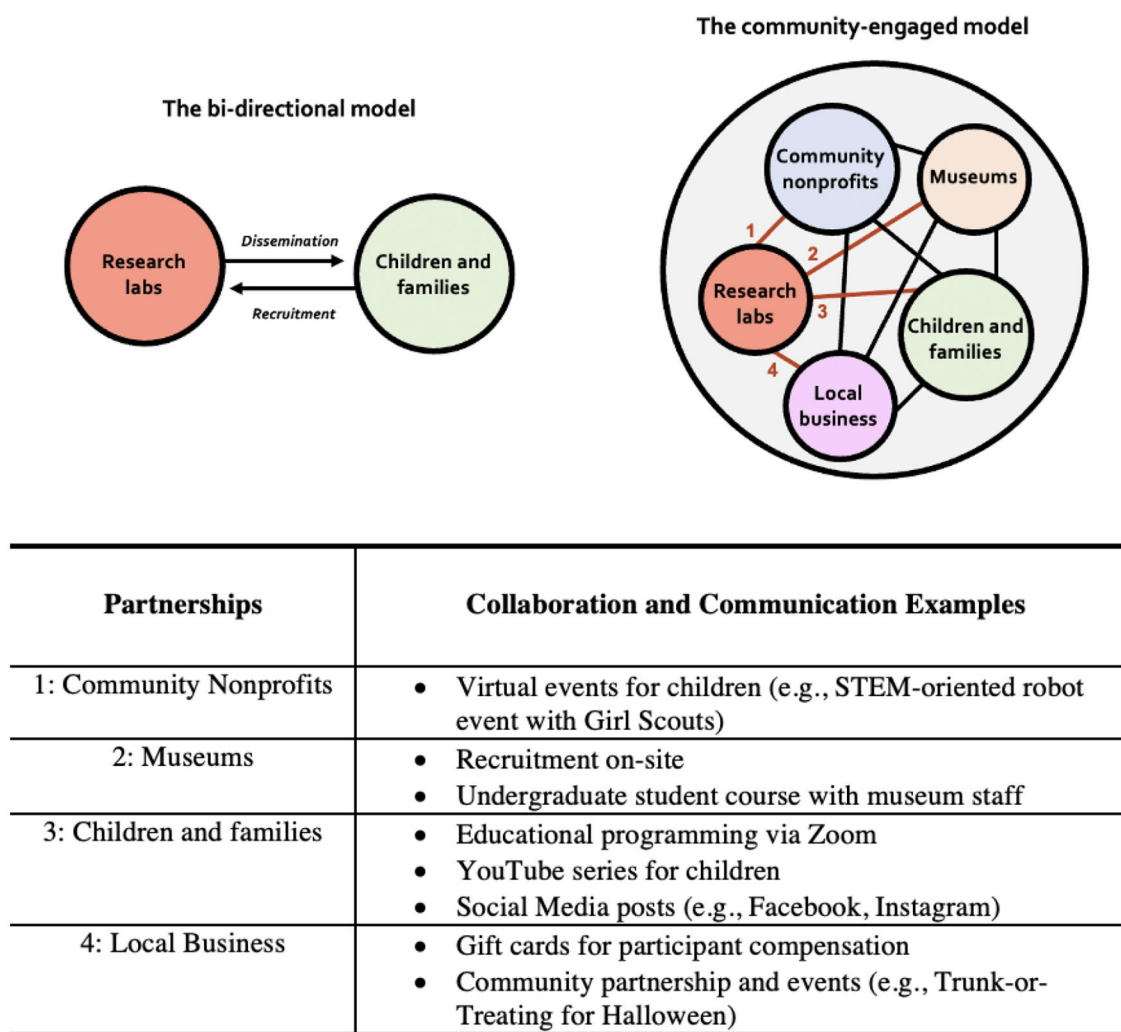


FIGURE 1 | The schematic depiction of the differences between the traditional model (left) and the community-engaged model (right) with a list of collaboration and communication examples (bottom). Under the traditional model, the typical relationship between developmental research labs and community stakeholders is a bi-directional exchange whereby communities provide data to scientists provide knowledge to communities. In the community-engaged model, research labs' explicitly articulate commitments to reinforce partnerships through resource and knowledge sharing based on reciprocity and mutual benefit (Kretzmann and McKnight, 1993).

lab identity, and to have a shared experience in the virtual world that resembled the one we had when we were physically together.

We chose to use the language of mission-driven organizations as a tool to help us stay connected. Throughout meetings, discussions of our lab *identity*—what it was, how it was changing, what we could do to maintain it in the face of massive change—were a motivating force driving us to keep going. Prior to the pandemic, we thought of ourselves as part of two communities: A local community of organizations serving children and families, and a global network of labs dedicated to developmental research. In our discussions early in the pandemic, we felt strongly that we wanted to emphasize our role in the local community even more, and we wanted to participate in children's lives as they were radically changing.

We also want to make clear that there is no “right” way to approach making these statements explicit. For us, a useful

starting point was to think about the research program of the lab, broadly construed, and to incorporate feedback from our closest community partners. We suggest beginning the process wherever it makes sense and being willing to see where it leads.

EMBEDDED IN A LOCAL COMMUNITY

Scientists are dedicated to creating knowledge, and often think of the infrastructure of their organization as merely a necessary means to this end. In this section, we examine the traditional lab practices involved in producing scholarship through the lens of a community-driven mission. **Figure 1** shows a visual representation of the *community-engaged lab* model side-by-side with the traditional, bi-directional model. The traditional model depicts a bi-directional relationship between developmental

science labs (which are members of universities and a broader network of scholars) and communities (either local or virtual) where children and families (our participants, who are also meant to be the beneficiaries of scientific knowledge) live. This bi-directional model is one that most developmental labs have in mind, at least implicitly, when they set up mechanisms for recruitment and dissemination. Our community-engaged model is different: developmental labs like ours and their scientific mission are *embedded* within a local community. Note that we do not mean this as a replacement for ways that labs connect with other communities, such as a global network of scholars, nor do we advocate operating separately from academic institutions in which labs reside. But rather, the community-engaged model's intent is to serve as a principled shift in guiding how we relate to local community organizations and the children and families which they serve. Instead of engaging with local communities from the outside, community-engaged developmental labs productively operate within, among, and alongside organizations that serve children and families in the local community.

Concretely, this model resulted in a reframing of many of our ordinary daily tasks—in particular, our public-facing activities. Recruitment and data collection are reframed under this model as *collaborative* actions: we grow our participant pool of child “citizen scientists” by supporting and sustaining long-term partnerships with other organizations whose mission, vision, and values align with ours. Dissemination of research findings is reframed as having the broader goal of science *communication*. In this way, all of our interactions with the community are opportunities for conveying our mission outwards. During the early days of the pandemic a top priority was to return, with minimal disruption, to our research. Our community partners—the science museum, local schools, and community centers which were our main avenues for reaching children and families prior to the shut-down—were experiencing their own upheavals. Our commitment to find new ways to stay engaged with the community impacted decisions we made as an organization about how to continue to work.

Below we describe how each reframing was put into action in our lab. For the purposes of illustration, the examples below are organized in two sections, with the recognition that the distinction is somewhat arbitrary. In the traditional bi-directional model, the dual-goals of recruitment and dissemination work together. In our community-engaged model, even as they are reframed, virtual collaborations and partnerships open up new opportunities for science communication. Efforts to engage in virtual science communication lead back toward goals of citizen science and toward samples of children that are more inclusive and representative of our local community.

Notably, the community-engaged lab model, at least in our case, was not intended as a change of direction away from our scientific mission, but rather a way of supporting it. Whether this approach has a long-term impact on our lab, on our trainees, or on our ability to do science, and whether it can be useful to other labs, remains to be seen. Despite this, we argue that for

the success of our own pandemic-era work, our identity as a community-engaged lab was critical.

RECRUITMENT REFRAMED

Under the traditional bi-directional model, recruitment *from* communities results in a supply of data *to* human-participant labs. One unintended consequence of this model of research participation is that it perpetuates the current predominance of homogeneous (predominantly white, predominantly high-SES) samples in research (Nielsen et al., 2017; Lourenco and Tasimi, 2020). Pre-pandemic, standard lab recruitment was successful for engaging with families that were already comfortable coming to labs, or those that had the time and resources to go to science and children's museums, or those that were familiar with (and trusting of) research protocols like signing consent forms. During the pandemic, these limitations were exacerbated by issues of availability of computers and stable internet connections, and the ability of parents to make time to schedule and connect through virtual lab visits—parents took on more roles as full-time caregivers, teachers for homeschooling, and, in some cases, employees working from home. Thus, it was no surprise that, initially, our lab (and perhaps others) saw samples becoming *less* representative of our local communities than they were before (see Figure 2).

We hoped that our community-engaged lab model could be a starting point for reaching groups of children and families that represent the diversity of our local community. In theory, this works by creating long-term sustainable collaborations with community partners, and establishing trust. But what about in the short term? Could this idea help us meet the immediate needs of functioning as a lab during the pandemic? Here we describe some examples, and signs of success.

Participant Incentives and New Community Partnerships

Under the bi-directional model, labs that have the means to provide incentives operate with the idea that gifts or other types of compensation are exchanges based on single interactions. If, for example, a family comes into the lab to participate in a study, they may leave with a gift or monetary compensation. If a school or museum partners with the lab, they too may receive gifts or donations.

Recognizing the indirect effects that our incentives have on communities, and reframing with this in mind, has led us to turn gifts and other forms of compensation into opportunities for more engagement with our local economy. For example, behavioral research labs often use Amazon gift cards to compensate participants for their time because Amazon is a globally accessible and convenient option. However, such an approach sacrifices a valuable opportunity to work with the local community, and an ethical duty to ensure that gift cards are used by children. To align with our community-engaged mission, when our lab transitioned to virtual studies, we sought out community partnerships with child-focused local businesses. Several factors motivated this

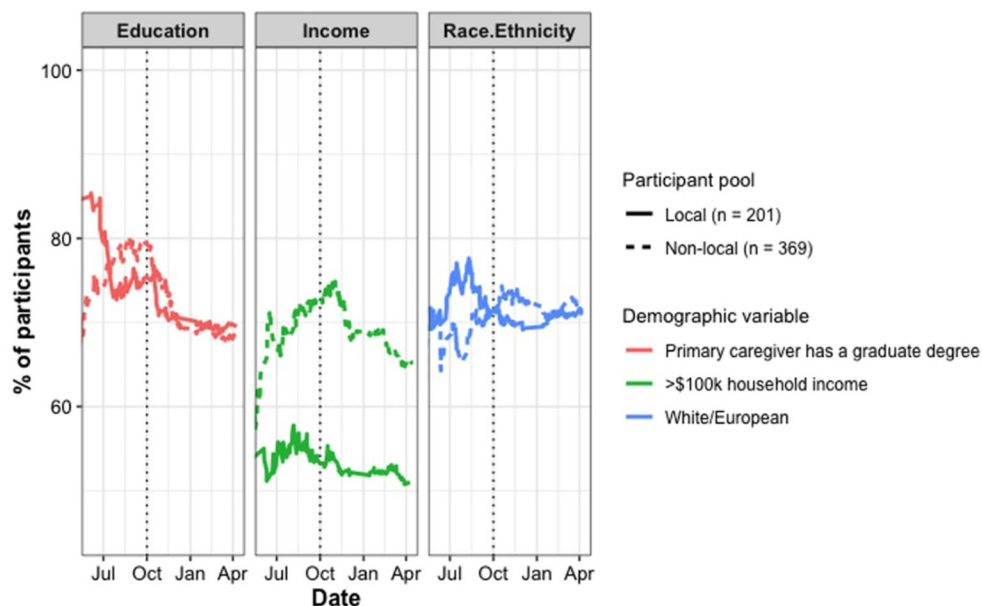


FIGURE 2 | Demographic indicators among local and non-local participants (June 2020–April 2021). The time series graph show the percentage of participants (starting $n = 30$ local, $n = 36$ non-local) whose primary caregiver has a graduate degree (left, red), whose annual household income is greater than \$100,000 (center, green), and who are white/European (right, blue). Solid lines indicate trends for local participants and dashed line indicate trends for non-local participants. High numbers on the y-axis indicate higher income, higher education, and predominantly white samples. The checked vertical line signifies the date of our lab meeting in October 2020, when we set new goals for reaching more participants outside of academic families.

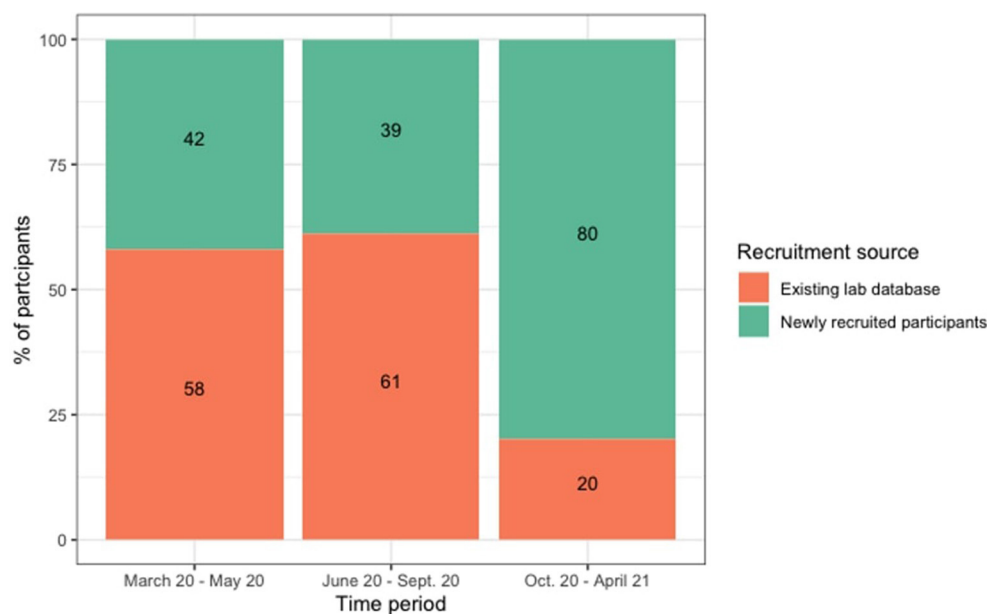


FIGURE 3 | Turnover from old to new participants over the course of the pandemic year. Percentage of participants from our pre-pandemic lab database (orange) and newly recruited participants (green) for each period of assessment (left: March 2020 to May 2020, center: June 2020 to September 2020, right October 2020 to April 2021).

search: we wanted the gift cards to be used for children, wanted to give participants alternative options to support local businesses, and wanted to start building collaborative community partnerships with organizations that were both a part of our

local economy and were mission-driven in the service of children and families.

We started by approaching a local toy store that was operating virtually via their website. In our initial meeting, we discussed our

common interests in children's learning. Our exchange resulted in a collaboration that benefited the business as well as our research: we created a mechanism for distributing gift cards that participants could use online or in-store (once in-person retail became an option again). Our lab handled the creation of advertisements (i.e., posters for the store) and the creation of visually appealing coloring-pages to serve as gift cards. We also trained researchers on how to offer the gift-cards as an alternative to Amazon without appearing coercive, similar to how we train researchers in informed consent. After a 6 month pilot program with the toy store, we expanded. We approached a local bookstore with a particular interest in children's education and literacy with the same idea, and tailored the gift-cards to this business, using their preferred system for keeping track of gift cards, and different advertising strategies.

Indicators of Success

It was informative for us to gather some data on the impact of our local business partnership programs. Of the 526 gift cards given out to all participants, local and remote, since we began online data collection, 429 (81.6%) requested Amazon gift cards and 97 (18.4%) requested local gift cards. Perhaps telling is that both local and remote participants requested local gift cards —21.6% (21/97) local gift cards went to participants outside of our county. Also telling is that local participants requested local gifts almost half the time —46% (76/165) requested local business gift cards rather than Amazon gift cards. We take this as a sign that many families were enthusiastic to support local businesses during the pandemic through our incentive program.

Of course, the goal of our relationship with both of these local businesses was not just to create alternative participant incentives, but also to build new community partnerships. Along these lines, we have maintained open discussions with each business with the idea that the collaboration could grow. Encouragingly, both business owners had multiple creative ideas about new ways to collaborate: the toy store suggested that we could help test out models of new toys that toy companies send to the store before they display it in store, and the bookstore was interested in hosting a book reading event for children. Although these events are planned, they are on hold until current pandemic restrictions are lifted, but we take the enthusiasm for continued collaboration as a concrete sign of success.

Museum Collaborations and Living Labs

Science and children's museums are mission-driven community organizations whose vision and values align well with developmental labs. There is already a decades-long tradition of developmental labs collaborating with local children's museums using "living labs" to recruit participants and collect data, and to disseminate findings on the museum floor (Sobel and Jipson, 2015; Callanan et al., 2020). The benefits of these partnerships have been noted before: for labs, museums offer convenient access to children and families, and a greater chance of reaching more representative participant samples though this sample is still limited to the patrons of local museums. For families, museum research reduces barriers associated with travel to

university labs, and increases opportunities to learn about developmental science and see it in action. For museums, the partnerships offer multiple benefits from positive visitor experience to opportunities for additional grant support.

Prior to the pandemic, our lab maintained such a partnership with our local science museum. We recruited and collected data on the museum floor in a living lab-style exhibit. We wrote and received several collaborative grants to support both organizations. We also designed and implemented several museum service-learning courses in which students worked on museum projects—ranging from designing exhibits and programs, to creating evaluation tools that could be used to measure impact.

The pandemic brought enormous challenges to museums, as they were forced to close to the public, furlough staff, and rethink their operations. From our perspective, it is tempting to view this as another example of how pre-pandemic ways of conducting research effectively shut down. But guided by our long-standing partnership, and by the community-engaged model that supported it, we take another view. On many fronts, members of our research team continued to engage with the museum and looked for ways to help support their work financially and logistically. In addition, the museum service-learning course ran virtually this spring, and included some in-person components. The museum project for students in spring 2021 was an evaluation of long-standing exhibits: students were assigned an exhibit to observe, conducted an evidence-based analysis of how the exhibit currently supports STEM learning, and, under the guidance of the instructor and museum staff, offered suggestions for improvements. Like its predecessor projects, this one was aligned with the training and mentorship missions of the lab, as well as the learning mission of the museum to create experiences for multi-age learning communities.

Community Events

In addition to growing long-term partnerships, we sought opportunities to participate in community events. On Halloween, our local mall hosted a Trunk-or-Treat event where families drove by a row of socially distanced vendors and local businesses to receive Halloween treats. A team of students from our lab planned for the event by preparing bags of treats with small flyers to advertise our studies to families. The event was a success, as we reached ~270 families who inquired about lab research and participation. Importantly, too, it was an opportunity for lab members to engage in outreach and develop their leadership and communication skills.

Single events can sometimes lead to long-term partnerships. For example, members of our lab connected with the local Girl Scout leaders to arrange a virtual STEM-oriented event about child-centered robot design. The initial idea was just a single virtual session: an educational program used to illustrate how humans and robots can "collaborate" on tasks. Over 40 children ages 4 to 13 participated in a virtual activity which asked them to direct a person pretending to be a "robot" around an obstacle course set up in their space. Children who had someone to collaborate with were provided instructions and examples of how

to give directions, and children who did not have someone helped direct a student “robot assistant” by giving directions over the conference call. The session concluded with an opportunity for Q&A about women in STEM and their career paths.

The event illustrates how the research and training goals of the lab can combine with addressing the needs of a community organization. From a research standpoint, the event was an adaptation of a recently published study which communicated the activity and research to the public. In addition, it exposed children to one of the major challenges in robotics: navigation in collaborative tasks. For training, it was an opportunity for our lab members to talk about their work and aspirations. For the community partner, they hoped to use the event to inspire and motivate girls in STEM careers. The troop leaders’ feedback after the event was overwhelmingly positive, and they have been very receptive and enthusiastic about future events that allow researchers to gather observational data and convey principles of design thinking to young girls.

DISSEMINATION REFRAMED

Under the traditional bi-directional model, dissemination usually lives in the space of academic discourse, such as publications and conferences. Most labs also include some dissemination to local communities—such as websites and lab newsletters—but this communication is generally thought of as completely separate. This accepted practice of scientific vs. public dissemination reinforces the separation between research and the communities it ultimately serves. For instance, it takes 17 years on average for findings in scholarly journals to reach the general public (Trochim, 2010). Additionally, if and when research findings do become more available, accessibility emerges as another barrier due to the financial cost of journal subscriptions, the time needed to thoroughly comprehend the studies, and the lack of readability as a result of the density of research jargon.

Adopting a community-engaged lab model encouraged us to think about lab *communication* rather than dissemination. Under this reframing, any opportunity for exchanges of knowledge with the community carries equal value. During the pandemic, this translated into actionable steps: we created internal mechanisms to train lab members to be good science communicators, and we used the tools of the virtual environment to make science accessible whenever possible.

Trainees as Community Ambassadors

Developmental labs commonly rely on young researchers-in-training to manage the many daily tasks involved in conducting human participant research with children and families. In the community-engaged lab model, trainees are not only a valuable resource to the lab, but a bridge between the lab and community in which it is embedded.

As the examples below illustrate, trainees in the community-engaged model become the most important resource for all public-facing communication activities. For us, it was important to ensure that each of these activities were free, inclusive, and accessible; we conceived them as low-stress ways for our lab members to interact with the public and encouraged each

trainee to think of themselves as an ambassador of science in the community.

A lot of ideas for activities were driven by the creativity of student trainees, who were responsive to community feedback: lab members often came back from virtual (or in-person) events with comments and suggestions from the children and families, and the mutual exchange of knowledge and ideas motivated further activities. Training students to think more about diversity in research, informing them on how a lab can serve the community, and encouraging them to take action has long-term benefits. While not all student trainees are interested in a career in research, many of them are interested in working with children and families across multiple professions. This experience as community-engaged lab ambassadors will translate to leadership and service in their future.

Exchanging Ideas for Communication and Outreach

To facilitate involvement of all lab members in communication and outreach, we devoted one meeting per month to discuss communication goals. The meetings were structured in the following way: At the beginning of the Fall 2020 semester, lab members chose to be part of one of three small teams: an *in-person team* that organized and participated in community events, an *outreach team* that focused on connecting with local community organizations, and a *social-media team* that advertised studies online to reach families outside of the local community. Each team nominated a student leader to ensure progress.

Acting in small teams (rather than as a larger group) leads to student trainees feeling heard in their ideas and having more ownership over their actions (Avey et al., 2009). Leaders and coordinators in the lab play an important role here: it is essential that these people genuinely understand the lab’s mission of promoting diversity and serving as a community engaged organization, and to constantly reflect on and assess the communication goals of the lab. When leaders in the lab actively communicate these values of promoting diversity and serving as a community engaged organization, it encourages conversations among lab members and establishes these values as the lab’s culture.

Communication Activities

Below we describe some of the results of this process, and give examples of our communication activities over the last year:

1. **Children Doing “Citizen Science.”** *Citizen science* is defined as “the collection and analysis of data relating to the natural world by members of the general public, typically as part of a collaborative project with professional scientists” (Oxford Languages, 2021). This idea motivates national recruitment efforts such as “Children Helping Science” which was organized by a consortium of labs in response to the pandemic shut-downs (Sheskin et al., 2020). Of course, the idea of citizen science is compatible with the bi-directional model as well. But under the community-engaged model, every time a child helps *us* as a “citizen scientist” is also an

opportunity for *us* to engage in science communication. This was particularly important this year, as families were stretched to their limits and children were spending more time in front of screens (Richtel, 2021). Consistent with our effort to re-imagine the lab, we emphasized to trainees and families that participating in research is not just a transaction, but rather an opportunity to learn about, and actively participate in doing science. Individual participant interactions, then, were bookended by discussion of the study purpose, opportunities to ask questions, and follow up newsletters with updates and findings, and other ways to get involved with the lab. To this end, 52% (186 out of 361) of children participated more than once in our online studies.

2. **Educational Programming via Zoom.** Over the summer, we hosted a series of educational programs designed for children between the ages of 3- and 8-years-old. Once a week, members of the lab took turns hosting 20–30 min sessions during which children would participate in an active-learning experience. All of the topics were based on lab members' own interests and ideas, with the only restriction being that it would be fun for preschool and school-age kids. Topics included visual illusions, robots, how to grow a garden, yoga, karate, origami, games, crafts, and more. Typically, anywhere between 3 and 10 children would join each week, and we found that children would commonly come back for repeated visits or for studies as a result. These programs offered children and families an informal, playful introduction to our team. They also kept all of us connected to children and to each other during the difficult summer when most other research engagement opportunities for students were unavailable.
3. **YouTube Kid's Series.** The fall semester brought new challenges for families facing school on screens. In our early fall reflections, and with feedback from parents, we recognized that a change to accommodate family schedules was needed. From this we moved our programming asynchronously, in the form of weekly 3–5-min YouTube videos. Once again, the focus was on active learning, play, and curiosity. Motivated by our own passions and interests, we wanted to inspire children to try something new, (e.g., learn a magic trick, make animations) or investigate a fascinating and perhaps unexplored phenomenon in the world around them (e.g., why do the leaves change colors in the fall?). In total, 22 kids joined us across 7 sessions during the summer. To date, the 17 videos in the YouTube series have received a total of 645 views.
4. **Social Media.** Like many other developmental labs, we use social media to reach potential participants locally and across the country. In addition to study advertisements, we have followed the growing trend of using social media for science communication. There are already successful campaigns on social media that directly aim to educate parents and practitioners about the science of child development. We viewed our efforts as an opportunity for research trainees to apply knowledge from the classroom to solving a real-world problem. As a group, we were encouraged to reflect critically about our science education: Was there a way to translate what we learn in labs and classrooms to community-engagement? We used reflective prompts, questions like:

“What is one thing you learned about child development that surprised you?” or “What research finding you’ve read inspires you?” We also talked about children’s lives and how they had changed, and looked for media and scientific coverage on the changing roles of parents. We aimed to keep our posts light, fun, and grounded in our own experiences. Our focus is on communicating curiosity, being ourselves, showing support for communities, rather than delivering information.

COMMUNITY ENGAGEMENT WORKS: MEASURING IMPACT

Throughout this paper, we have emphasized how a simple shift in perspective—thinking of developmental labs as embedded within a network of local community organizations—can help engender a number of positive outcomes for local children, families, community partners, and early career trainees. Throughout this paper, we have shared anecdotal evidence of such impacts: we created two new, long-lasting community partnerships, we hosted educational events and weekly programming that together encouraged hundreds of local children to be active, curious “citizen scientists.” All along the way, early career trainees played a crucial role in fostering such relationships, and in turn gained valuable leadership skills.

In addition to this anecdotal evidence of impact, we also have internal data that shows the community-engaged lab model’s role in making our online participant pool quantifiably more inclusive and representative of our local community. Next we discuss how our lab assessed and modified our community-engaged recruitment aims by analyzing the standard demographic information collected from our study consent forms.

Did our community engagement efforts have a measurable impact on the demographics of our study participants? To assess this, we looked at how the percentage of local participants (In Ithaca, NY and nearby area codes) who were from highly educated (caregiver has a graduate degree), high annual income (>\$100,000), and White/European households compared to the respective levels from the previous calendar year (February 2019–March 2020). For a summary of these findings, see **Table 1**, **Figures 2, 3**. Below, we provide case study details that illustrate how we used these data to help inform our engagement efforts throughout the year.

At the outset of online data collection (March 2020–April 2020), all three of the indicators had increased above their pre-pandemic baselines ($n = 30$ collected online; education: +18%, income +12%, white/European: +4%), confirming our lab’s shared sense that data collection had become more narrowly confined to academic social networks. In the early Summer of 2020, our discussions about community engagement and explicitly acknowledging our mission-driven approach set us on the course of expanding our outreach efforts. As reviewed earlier, we took on several engagement initiatives: partnering

TABLE 1 | Quarterly summary of demographic indicators.

| Date | Education Primary caregiver has a graduate degree | Income Annual household income > \$100k | Race/Ethnicity White/European |
|---------------------------------------|--|--|----------------------------------|
| Pre-pandemic (Feb. 2019–Mar. 2020) | 66% | 51% | 77% |
| Spring '20 (April 2020) | 84% | 63% | 81% |
| Fall '20 (October 2020) | 77% | 63% | 71% |
| Spring '21 (April 2021) | 69% | 60% | 71% |

A summary of the percentage of participants whose primary caregiver has a graduate degree ("Education" column), whose annual household income is >\$100,000 ("Income" column), and who are white/European ("Race/ethnicity" column). Each row summarizes these indicators at our periodic assessment dates: a pre-pandemic baseline (February 2019–March 2020) and once per semester to date (April 2020, October 2020, April 2021).

with local businesses to offer gift cards, hosting free educational programming for children, and more.

In early October 2020, we met again as a lab to discuss how the most recent (April 2020 - October 2020) demographic data compared to the pre-pandemic baselines. We found that, in general, our participant pool had indeed become more representative of our community since April 2020, in the initial lock-down ($n = 215$: education: -7% , income -0% , white/European: -10%), but the education and income figures in particular were not yet back to the same level as pre-pandemic (overall, since February 2020, $n = 245$: education: $+11\%$, income $+12\%$, white/European: -6%). With this in mind, we dedicated time at our weekly lab meetings for targeted discussions about reaching out to more children and families from non-academic and lower-income households, as well as continuing to make strides in reaching a more racially and ethnically diverse group of children. From these conversations sprung many of our community-centered initiatives: collaboration with the local bookstore, a more consistent YouTube series, the Girl Scouts event, and more.

In the Spring of 2021, we met again to assess our lab's progress via the same indicators as before. The time series data (Figure 2) showed a consistent trend of our lab reaching more participants who do not come from academic families nor families in the top income bracket on our consent form (since October 2020, $n = 325$: education: -8% , income -3% , white/European: -0%). Indeed, a year into online data collection, all three indicators are trending toward (or back to) the pre-pandemic baselines for our lab (since April 2020, $n = 540$: education: -15% , income -3% , white/European: -10% ; overall, since March 2020, $n = 570$: education: $+3\%$, income $+9\%$, white/European: -6%).

Though sampling from a diversity of communities is important in its own right, it is equally important to have some objective measures to compare our analysis with the overall demographics of our local and non-local participants. For this we drew from current US Census data (U. S. Census Bureau, 2019). Our local county is 77.1% White, 29% of households have an annual income > \$100k, 33.1% have a

postgraduate degree. Overall the entire US is 60.7% White, 31.4% of households have an annual income > \$100k, 12.8% have a postgraduate degree.

A comparison suggests that our local participant pool is representative of the racial/ethnic make-up of our local community (71.6% white vs. 77.1% census baseline), whereas the non-local participant pool disproportionately samples from white populations in comparison to the national average (70.7% white vs. 60.7%). However, both our local and non-local participant pools disproportionately sample from households with higher income (local: 50.7 vs. 29%, non-local: 65.3 vs. 31.4%) and higher educational attainment (local: 70 vs. 33.1%, non-local: 68.6 vs. 12.8%). However, our local participant pool is comparatively much more representative, as it is $\sim 20\%$ closer to the census baseline on both indicators.

This analysis was helpful and informative for us as a measurable target for assessing whether progress was made. Ultimately, the most meaningful measures of community engagement will depend on the particulars of the lab and their local community. We share these data to illustrate how our community-engagement efforts led to quantifiable impacts on the representativeness of our lab's subject pool and suggest that labs can tailor measures of impact to their own communities.

CONCLUSION

How has the pandemic changed developmental research? On the surface, it has resulted in a slew of new challenges including a sudden transition to online data collection, an abrupt discontinuation of on-going studies, and the loss of access to physical lab spaces. In addition, the pandemic has made existing challenges newly visible (Benner and Mistry, 2020; Yip, 2020; Sonnenschein et al., 2021). A new perspective on our organizations is exactly what we need, both for the current times and for the transition to in-person work in the future.

Here we have tried to make a case for adapting the ideas of mission-driven community organizations and showed that this approach was critical to our success in a difficult year. We began with establishing explicit statements of *mission, vision, and values*, and used them to start internal discussions toward developing a "community-engaged" lab identity, acknowledging that our developmental lab is *embedded* in a community of like-minded organizations working on behalf of children and families. We also demonstrated how this approach allowed us to adapt to our changing circumstances. Our community-engaged mission guided our decisions about even the most ordinary lab tasks, such as recruitment, data collection, dissemination, and involvement in training and mentorship of students.

We hope to see other examples of labs develop their own broader vision that goes beyond the standard research and training missions common to labs like ours. The metaphorical and physical partition between the community and developmental labs inside academic institutions perpetuates many of the gaps between research and practice that developmental scientists are all too familiar with. It has long been recognized that our traditional bi-directional

exchanges perpetuate homogenous samples and thus limited generalizability, little dissemination of research findings outside of academia, and research topics that often do not appreciate the assets and address the needs of educators and practitioners. Like others, we would like to see these barriers lifted. Furthermore, we believe embedded, community-engaged labs also contribute to a more positive public perception of science.

There is no “right” way to start this process. From our experience, simple actions are the best place to start, and local needs serve as a guide. Over the last year, we started by expanding to new partnerships, maintaining our existing partnerships while remaining sensitive to their pandemic-era needs, facilitating live and pre-recorded forms of educational programming, sharing newsletters, and increased social media presence. Community relationships are further strengthened when developmental labs are intentional about creating positive and meaningful interactions with children and families in *every* session.

We further believe success critically depends on empowering trainees in their dual role: they are not only researchers, but also community ambassadors. As part of mentorship and training, it is standard practice in our lab to teach students to be a resource to the community. In this way, students learn valuable skills that translate to work outside of the lab and classroom, such as leadership, communication, ethical practice, and how to work for social justice and change.

Adopting this perspective does not have to influence the kind of research one does, but it can. For example, volunteering at a local science museum might cause one to notice that some groups of children are more likely to participate in certain events than others, prompting follow up questions of why that is and how to change it. In the same way, a researcher hosting a virtual live educational program with children might wonder how the pandemic is impacting the way children think, learn, and feel about the people and the world around them. In fact, a research project started by several members of our lab grew out of our experiences engaging with children and families over the summer of 2020. We were inspired by conversations with children during our online programs to add to a growing number of studies on the topic of children’s psychological well-being during the many pandemic-era transitions (Medlin, 2000; Laursen et al., 2007; Sun et al., 2020; Tso et al., 2020). Because our emphasis included building collaborative relationships with individual families, we were able to follow up with the same children to track longitudinal change in well-being over the course of the year. In sum, with this shift in perspective, labs can continue to do the research they were initially passionate about and stay open to new ideas that respond to changing needs and current events.

But research in a community engaged lab needs to always happen in the community, or outside the lab. There is a difference between adopting a community-engaged lab model as a guiding principle to *run a lab organization* (i.e., explicitly stating mission, vision and core values, viewing the lab as “embedded” within a community of organizations that care about children and families) and doing *community engaged research*. The distinction is critical: not all community engaged labs do community engaged research. Some (ours included) do basic research, and some of that work has to be done in the lab under certain

conditions. But even basic research labs can openly care about how we connect and engage with our communities and devote some of our time and efforts to doing so.

We take our ability to adapt to changing circumstances and continue to conduct research as signs of success. But the benefits of our approach came in many other forms as well. Through our discussions, we maintained a sense of connection to each other despite physical isolation. We formed relationships with new participating families and new organizations in the community. We helped support the local economy. We were able, after less than a year, to return to pre-pandemic levels of demographic representation.

Of course, none of these measures of impact are an endpoint. For one thing, we do not yet have evidence that this approach does a *better* job than the traditional model—or the newer online platforms that encourage broad participation nationally and internationally—in reaching children from backgrounds currently underrepresented in developmental science. In our view, investment in local communities works together with these national efforts toward more inclusive scientific practice. For one thing, the more that labs embed themselves within their local communities, the more they can meaningfully contribute to multi-site collaborations in a broader network of scholars (e.g., Frank et al., 2017). Thus, we believe that local engagement can be a mechanism for diversifying our field.

The process we present here was not without its challenges. Re-imagining lab identity requires an enormous up-front cost in time and resources that could be spent in other ways. We therefore recognize that this level of investment perhaps could only have happened in an extraordinary year, when many other activities were impossible. Quite frankly, our lab benefitted from the lack of other jobs and internships seeking to employ undergraduate students in the summer of 2020. Everyone stayed (which is not typical) and thus our work could continue over the summer months. It was unusual even for us to have our full staff of researcher trainees volunteering all year including summer, and for other smaller labs with fewer undergraduate researchers (and perhaps little or no graduate students) the picture will look very different. We devoted many hours to discussions of community engagement—time that could also be devoted to reading scientific journal articles, presenting our work for feedback, and other discussions. Admittedly, we do not have data on the number of person-hours (at various career stages, including PI and graduate student hours) that it takes to setting up a community-engaged lab at the expense of other work. We do however want to note that all of us were also teaching (online this year) and maintaining our administrative roles within the university, but of course these non-research activities vary significantly from one university to another. We therefore acknowledge that lab size, lab resources, time, and funding may be limiting factors. For this reason, we use our year by way of example only, and caution against creating a set of recommendations that are suitable for all.

Where do we go from here? We have tried to show that a year of community-engagement can yield measurable benefits, but we do not yet know how this will affect the transition back to in-person work. We expect that over the coming year reopening

in-person labs will present new challenges, as will reopening of schools, museums, community centers, and other spaces which play a role in children's lives. Perhaps the last lesson we take from this experience is that the world is constantly changing, and if we act in ways that are responsive to change, we will, as scientists, get closer to understanding children in the ecologies in which they develop.

ADDITIONAL READINGS/RESOURCES

For more examples of high-impact mission-driven organizations:

- Crutchfield, L. R., and Grant, H. (2007). *Forces for Good: The Six Practices of High Impact Nonprofits*. San Francisco, CA: Jossey-Bass.

For more readings on modeling university-community partnerships:

- Kretzmann, J. P., & McKnight, J. L. (1993). *Building Communities from the Inside Out: A Path Toward Finding and Mobilizing a Community's Assets*. ACTA Publications.
- Asset-Based Community Development (ABCD) Institute. (2021). *ABCD Institute*. <https://resources.depaul.edu/abcd-institute/Pages/default.aspx>.
- There are many examples of applications of ABCD to organizations that serve children and families, including libraries and museums:
 - Baron, D. (2020, November 25). *Libraries and Museums as Catalysts for Change*. Steans Center. <https://resources.depaul.edu/steans-center-community-based-service-learning/about/news/Pages/Libraries-and-Museums-as-Catalysts-for-Change.aspx>.

REFERENCES

- Ash, S. L., and Clayton, P. H. (2009). Generating, deepening, and documenting learning: the power of critical reflection for applied learning. *J. Appl. Learn. Higher Educ.* 1, 25–28.
- Avey, J. B., Avolio, B. J., Crossley, C. R., and Luthans, F. (2009). Psychological ownership: theoretical extensions, measurement, and relation to work outcomes. *J. Organ. Behav.* 30, 173–191. doi: 10.1002/job.583
- Benner, A. D., and Mistry, R. S. (2020). Child development during the COVID-19 pandemic through a life course theory lens. *Child Dev. Perspect.* 14, 236–243. doi: 10.1111/cdep.12387
- Callanan, M. A., Legare, C. H., Sobel, D. M., Jaeger, G. J., Letourneau, S., McHugh, S. R., et al. (2020). Exploration, explanation, and parent-child interaction in museums. *Monogr. Soc. Res. Child Dev.* 85, 7–137. doi: 10.1111/mono.12412
- Crutchfield, L. R., and Grant, H. (2007). *Forces for Good: The Six Practices of High Impact Nonprofits*. San Francisco, CA: Jossey-Bass.
- Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., et al. (2017). A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy* 22, 421–435. doi: 10.1111/inf.12182
- Kretzmann, J. P., and McKnight, J. L. (1993). *Introduction to Building Communities from the Inside Out: A Path Toward Finding and Mobilizing a Community's Assets*. Chicago, IL: ACTA Publications.
- Laursen, B., Bukowski, W. M., Aunola, K., and Nurmi, J. E. (2007). Friendship moderates prospective associations between social isolation

and adjustment problems in young children. *Child Dev.* 78, 1395–1404. doi: 10.1111/j.1467-8624.2007.01072.x

- Ash, S. L., & Clayton, P. H. (2009). Generating, deepening, and documenting learning: The power of critical reflection for applied learning. *Journal of Applied Learning in Higher Education*, 1(1) 25–28.
- Kiely, R. (2015, October 13). *Considering Critical Reflection*. Global SL Blog. <https://compact.org/criticalreflection/>.

AUTHOR CONTRIBUTIONS

JL, SP, YS, and TK contributed to conception and organization of the paper. SP performed the data analysis and created data visualizations. JL, SP, and YS wrote the first draft of the manuscript with revisions from TK. MK and RK contributed to sections related to research-practitioner partnerships. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the National Science Foundation DS 1823658 grant and National Institute of Food and Agriculture Hatch grant NYC-321434 to TK.

ACKNOWLEDGMENTS

Thanks to our new community partners Greta Perl and Laura Larson. Thanks to members of the Cornell Early Childhood Cognition Lab Marianella Casasola and Melissa Koenig for helpful comments on earlier versions of the manuscript.

- and adjustment problems in young children. *Child Dev.* 78, 1395–1404. doi: 10.1111/j.1467-8624.2007.01072.x
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Medlin, R. G. (2000). Homeschooling and the question of socialization. *Peabody J. Educ.* 75, 107–123. doi: 10.1207/S15327930PJE751&2_7
- Nielsen, M., Haun, D., Kärtner, J., and Legare, C. H. (2017). The persistent sampling bias in developmental psychology: a call to action. *J. Exp. Child Psychol.* 162, 31–38. doi: 10.1016/j.jecp.2017.04.017
- Oxford Languages (2021). In *Oxford English Dictionary*. Retrieved from: <https://www.oed.com/browse/dictionary> (accessed July 22, 2021).
- Richtel, M. (2021). *Children's Screen Time has Soared in the Pandemic, Alarming Parents and Researchers*. The New York Times. Available online at: <https://www.nytimes.com/2021/01/16/health/covid-kids-tech-use.html>
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Sobel, D. M., and Jipson, J. L. (2015). *Cognitive Development in Museum Settings: Relating Research and Practice*. New York, NY: Routledge. doi: 10.4324/9781315667553
- Sonnenschein, S., Stites, M., and Ross, A. (2021). Home learning environments for young children in the U.S. during COVID-19. *Early Educ. Dev.* 32, 794–811. doi: 10.1080/10409289.2021.1943282
- Sun, Y., Li, Y., Bao, Y., Meng, S., Sun, Y., Schumann, S., et al. (2020). Brief Report: increased addictive internet and substance use behavior during the

- COVID-19 pandemic in China. *Am. J. Addict.* 29, 268–270. doi: 10.1111/ajad.13066
- Trochim, W. (2010). “Translation won’t happen without dissemination and implementation: Some measurement and evaluation issues,” *3rd Annual Conference on the Science of Dissemination and Implementation*, 581–629.
- Tso, W. W. Y., Wong, R. S., Tung, K. T. S., Rao, N., Fu, K. W., Yam, J. C. S., et al. (2020). Vulnerability and resilience in children during the COVID-19 pandemic. *Europ. Child Adolesc. Psychiatry*, 1–16. doi: 10.1007/s00787-020-01680-8
- U. S. Census Bureau (2019). *American Community Survey 1-year estimates*. Census Reporter Profile page for Tompkins County, NY. Retrieved from: <http://censusreporter.org/profiles/05000US36109-tompkins-county-ny/> (accessed July 26, 2021).
- Yip, T. (2020). *Statement of Evidence: Addressing Inequities in Education During the COVID-19 Pandemic: How Education Policy and Schools Can Support Historically and Currently Marginalized Children and Youth* (Policy Brief). Retrieved from: <https://www.srcd.org/research/addressing-inequities-education-during-covid-19-pandemic-how-education-policy-and-schools> (accessed July 26, 2021).
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Liu, Partington, Suh, Finiasz, Flanagan, Kocher, Kiely, Kortenaar and Kushnir. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Children's (Mis)understanding of the Balance Beam (Online Edition)

Virginie M. L. Filion* and Sylvain Sirois

Groupe de recherche sur la cognition, les neurosciences, l'affect et le comportement (CogNAC) and Département de Psychologie, Université du Québec à Trois-Rivières, Trois-Rivières, QC, Canada

The balance-scale task, proposed by Inhelder and Piaget, illustrates children understanding of weight-distance relationships. Piaget used the clinical interview method in order to investigate children's reasoning. Over the last five decades, Siegler's Rule-Assessment Approach has been used to explain children reasoning in the balance-scale task according to rules children would use to solve the task. However, this approach does not take into account some key perceptual properties of the task. This study evaluates whether different task demands would alter children's errors. Forty children (twenty children aged 4–5 years and twenty children aged 9–10 years) predicted the movement of both arms of 16 balance-scale problems administered online. Nine 4–5-year-olds produced non-plausible responses whereas none of the 9–10-year-olds provided non-plausible responses. These results seem to indicate a basic misunderstanding of the scale from some younger children, one that eludes traditional measures used with this task.

OPEN ACCESS

Edited by:

Rhodri Cusack,
Trinity College Institute of
Neuroscience, Ireland

Reviewed by:

M. Teresa Anguera,
University of Barcelona, Spain
Amy Masnick,
Hofstra University, United States

*Correspondence:

Virginie M. L. Filion
virginie.maude.laverdure@uqtr.ca

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 29 April 2021

Accepted: 30 July 2021

Published: 23 August 2021

Citation:

Filion VML and Sirois S (2021)
Children's (Mis)understanding of the
Balance Beam (Online Edition).
Front. Psychol. 12:702524.
doi: 10.3389/fpsyg.2021.702524

Keywords: cognitive development, children, balance-scale task, online testing, rule-assessment, clinical interview

INTRODUCTION

The balance-scale task (Inhelder and Piaget, 1958) is a logicomathematical problem-solving task. The scale consists of two arms in the form of a unitary beam, centrally attached to a fulcrum. On each arm, there are pegs placed at equally spaced distance from the fulcrum which are used to place unit weights. The child's task is to predict whether the left or the right arm will tilt down, or whether the unitary beam will remain in balance. Children's understanding of the weight-distance relationship with force (i.e., the torque applied to the arms) is examined according to their responses.

There are six typical problems used with the balance-scale task (Siegler, 1976). These problems manipulate the weight-distance relationship in different ways. There are three non-conflict problems (balance, weight, and distance) and three conflict problems (conflict-weight, conflict-distance and conflict-balance). In non-conflict problems, at most one parameter (weight or distance) differs on both arms. For weight problems, the weight values differ but distance is the same on each side of the fulcrum. For distance problems, weights are equal on each side but distances differ. In weight and distance problems, the side with relatively largest value tilts down. For balance problems, the values of weight and distance are identical on each arm and the beam remains stable. In conflict problems, the weight and the distance values differ on each arm of the scale. For conflict-weight problems, the arm with relatively more weight creates relatively more torque and tilts down. In conflict-distance problems, it is the arm with the relatively larger distance from the fulcrum that tilts down. Finally, in conflict-balance problems, the combination of weight and distance on each arm creates the same torque and the beam remains stable (Siegler, 1976; Halford et al., 2002).

In line with the general idea of sequential stages of development (Piaget, 2002), it was initially believed that children go through three stages of development in order to solve the task (Inhelder and Piaget, 1958). It was argued that around 5–8 years of age, children acquire an understanding that their actions can impact those of an object. Children thus begin to understand the impact of weight and distance on the scale. However, 5–8-year-olds do not seem to be able to successfully combine the values of weight and distance together. This coordination of information would be understood around adolescence (Inhelder and Piaget, 1958).

The mathematical solution to solve the balance-scale task is to calculate torque. The torque, product of weight and distance, represents the force applied to one side of the scale (Inhelder and Piaget, 1958; Ferretti and Butterfield, 1992; Shultz et al., 1994). The arm with the largest torque will be the one that tilts down. When torques are equal, the beam remains balanced.

Siegler (2016) suggested that development is more like an overlapping wave model. A child could have different strategies (with variable probabilities of use) available at any given time (Siegler, 2016). Siegler and Chen (1998) explain development as a continuum where children dynamically add and select increasingly complex rules. The Rule-Assessment Approach (Siegler and Richards, 1979; Siegler and Chen, 2002) explains that children solve the balance-scale task according to four rules depending on their understanding of the weight-distance relation. Children who have no understanding of the weight-distance relationship would solve the task by chance (Rule 0). They would have an average success rate of 33% on any given trial since they have three answer choices available (i.e., balance, left, or right; Siegler, 1976). Over development, children begin paying attention to weight (Rule 1), then later consider distance when weights are equal (Rule 2), then try and fail to integrate both dimensions (Rule 3), until they successfully compare torques (Rule 4; Siegler, 1976).

Over the years, Siegler's work has been criticized because those four rules explain the reasoning of only 88% of children (Zimmerman, 1999). Jansen and van der Maas (2002) suggested that children can use multiple other rules. There are additive rules, multiplication rules, and perceptual rules (Ferretti and Butterfield, 1986; Jansen and van der Maas, 1997, 2002; Richardson et al., 2006; Messer et al., 2008; Hofman et al., 2015).

The balance-scale task can be an intuitive task if children rely on perception to solve the task (Shultz and Takane, 2007). The torque effect is a perceptual effect caused by the relative salience created on each side of the scale. A bigger difference between the torque of each side of the scale makes it easier for the child to solve the problem (Ferretti and Butterfield, 1986; Hofman et al., 2015).

Task demands could also have an impact on performance (Messer et al., 2008; Hofman et al., 2015). One study examined 4–5-year-olds' basic understanding of the task (Sirois et al., 2005). Using computer-generated images and videos of a balance-scale task where only weight was manipulated (distance was constant on all problems), the authors found that children did not understand the unitary nature of the beam in the apparatus, and given the opportunity would predict impossible behavior

from the balance (e.g., both arms down). Published studies use methods that only invite plausible answer choices (Siegler, 1976; Ferretti and Butterfield, 1986; Halford et al., 2002; Hofman et al., 2015). Indeed, recent studies used artificial neural network models (Zon and Xie, 2014; Shultz, 2017; Al-Atrash et al., 2020) to replicate findings of rule-assessment. Children's basic understanding assumption (i.e., the unified character of the scale) remains unchallenged. There is a real possibility that the bulk of the literature on this task has either overestimated children's performance, and/or mischaracterized their errors.

The main objective of this study is to evaluate whether different task demands would reveal different errors. Specifically, we predict that younger children (aged 4–5) do not understand the unified character of the scale (Sirois et al., 2005). Therefore, a proportion of errors will stem from predicting impossible behavior of the scale.

With a different methodology, it is unclear whether the torque effect would remain beneficial, or further compound the misunderstanding of the scale for younger children. Therefore, we manipulate the relative torque across problems, but only predict a beneficial effect for older children.

Finally, for exploratory purposes, we introduced a salient feature to help children focus on the dynamic aspects of the balance-scale, and not just static states. A bell was randomly placed above or below the scale for each child, to create a shift of focus from end states (L, R, or balance) to transformations (upward or downward motion). We predict that this salient feature will affect the types of impossible answers of younger children, given their purported relatively simpler understanding of the scale, if they are nevertheless sensitive to transformations (Sirois et al., 2005). A bell below is expected to enhance their implicit use of torque, whereas a bell above should disrupt it. In both cases, it may provide a finer-grained interpretation of their understanding of the scale.

METHOD

Participants

Forty children participated in the experiment: twenty 4–5-year-olds (13 girls and 7 boys; mean age = 61.1 months, SD = 7.49) and twenty 9–10-year-olds (8 girls and 12 boys; mean age = 116.9 months, SD = 6.09). No child had a diagnosis of learning or developmental disability, and all had normal eyesight and hearing. Children were recruited through Facebook pages that reach parents in various cities of Québec, Canada. All parents had to provide written consent for their child to participate in the experiment. This experiment was approved by the Comité d'éthique en recherche avec des êtres humains de l'Université du Québec à Trois-Rivières.

Materials and Stimuli

Scale

A wooden (Figure 1), purpose-built scale 27-inches high and 20-inches wide was used. Each arm of the balance was 10-inches and had six pegs on the top and one on the bottom. Pegs were 1.5-inches apart. The right side and its first five pegs from the fulcrum were red. The left side and its first five pegs were blue.

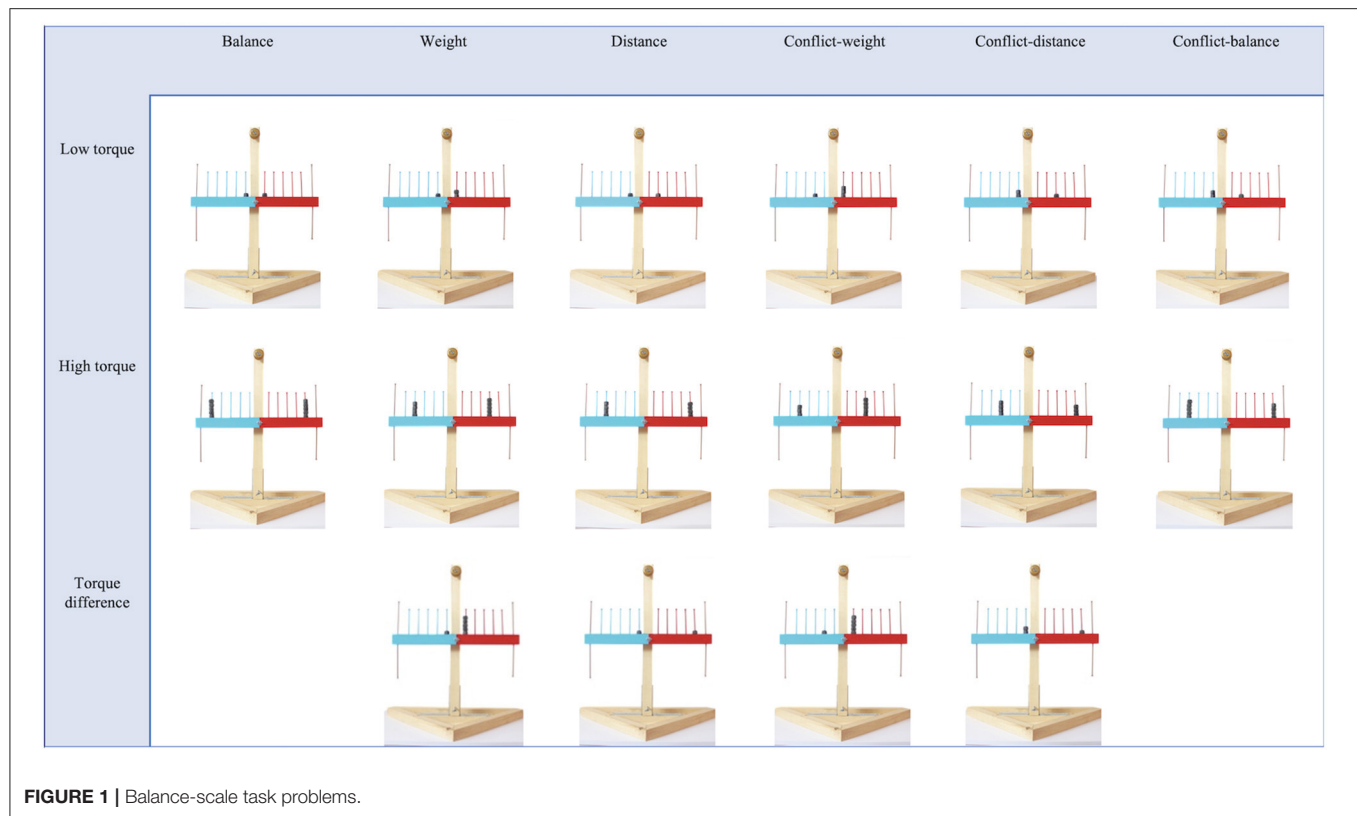


FIGURE 1 | Balance-scale task problems.

Red and blue pegs were 4-inch high. They were used to place the weights. The four pegs furthest from the fulcrum were brown and 5.5-inches high. They were used to ring a golden bell placed above or below the scale. This bell was at a distance of 11-inches from the fulcrum. As the arms of the scale are one united piece of wood, any difference in torque between left and right arms would cause a single, unified motion (left-down/right-up or left-up/right-down). The weights were hexagonal metal nuts, 0.8-inch in circumference, 0.5-inch high and painted black. These nuts weighed 18 grams. A maximum of five nuts could be put on each peg.

Stimuli

Thirty pictures of the scale with different weight-distance configurations on each arm were taken on a white background. There were 16 different types of problems (see examples in **Figure 1**), and each problem where the weight-distance configuration was different for both arms was duplicated to counterbalance left and right combinations. For problems with alternative images, one of the images was randomly selected for each child. Three types of torque were used in this experiment. The low torque (LT) problems had small weight and distance values on each side of the scale. One trial of each problem types (weight, distance, balance, conflict-weight, conflict-distance, and conflict-balance) was presented. The torque difference (TD) problems had one side with small weight and distance values, the other with large weight and/or distance, creating a large difference between both sides. Problem types of weight, distance,

conflict-weight and conflict-distance were used with TD (no balance problems can be created for TD trials). Finally, the high torque (HT) problems had large weight and distance values on each arm. Torque differences were also larger for the HT problems than for corresponding LT problems. Each of the six problem types were presented in the HT subset. The order of the 16 problems presented to the children was counterbalanced and randomized. Bell position was randomized between children.

Procedure

The experiment was a 15-min meeting on the videoconferencing platform Zoom. A script was developed to standardize the procedure across participants. The experiment began with a presentation of the scale. The researcher showed the position of the bell to the child. A short demonstration allowed the child to hear the sound of the bell when the scale tilted either side. Then, in the manipulation phase, the child could see five simple movements of the scale. Each time children saw the movement, they had to explain what the balance did. After the demonstration, the child was invited to choose three weight-distance configurations of their choosing and test the scale behavior for each.

Then, 16 pictures of the scale were sequentially presented to the child, who had to predict the movement for each side of the scale. For each picture, the child was asked “what does the blue side do” and “what does the red side do.” The question order (red then blue, or blue then red) was counterbalanced and randomized for each child.

To keep children engaged, there were seven predetermined encouragements during the experiment (e.g., “good,” “you are doing fine”). Times of encouragements (after trials 1, 4, 8, 10, 12, 14, and 16) were chosen randomly. They were independent of performance, so should not introduce systematic biases.

Children could take breaks if needed or stop the experiment at any time. Both age groups had the same procedure. Children were allowed to change their answers when they considered they made a mistake on their first attempt. Their second answer was, then, used for the analysis. At the end of the experiment, the child and parent were thanked by the researcher.

Data Preprocessing

Raw data were compiled using Matlab. Performance on each trial was scored 1 when correct, 0 otherwise. Non-plausible answers are erroneous responses whereby children predicted a violation of the rigid and unitary nature of the arm. They were coded as “BothDown,” “LeftDown,” “RightDown,” “BothUp,” “LeftUp,” and “RightUp.” Responses coded as “BothDown” involve a prediction of both arms down. The code “LeftDown” means the child predicted the left arm went down and the right arm remained stable. For “LeftUp,” the child would have predicted that the left arm went up and the right remained stable. All implausible errors involving downward motion were tallied into “TotalDown” scores; those related to upward motion were tallied into “TotalUp” scores.

Children were also classified according to Siegler’s rules. Rules 0–4 create unique sets of predictions (correct, wrong, guess) for each of the 16 problems. Using Euclidean distance (e.g., Aldenderfer and Blashfield, 1984), the square average distance between children’s performance on all 16 trials (1 for correct, 0 for wrong) and the predictions from each rule for those trials (1, 0, and 0.333 respectively for correct, wrong, and guess) were computed. Children were assigned the rule associated with the least Euclidean distance from their performance (see **Supplementary File** for details).

RESULTS

Out of 16 problems, the mean number of correct answers was 8.5 (95%CI [7.42; 9.57]). Younger children (4–5-year-olds) had an average of 6.05 correct answers (95%CI [4.93; 7.17]). Older children (9–10-year-olds) had an average of 10.95 correct answers (95%CI [9.91; 11.99]). An independent-sample *t*-test revealed a significant difference between the two age groups, $t(38) = -6.72$, $p < 0.001$, Cohen’s $d = -2.13$.

Table 1 shows children classification according to Siegler’s rules. A Chi-square test of independence indicated a significant association between Siegler’s rules and children’s age [$X^2(4) = 17.28$, $p = 0.002$, $V = 0.66$].

Younger children (aged 4–5) produced 41 non-plausible responses whereas 9–10-year-olds did not provide non-plausible responses. A Chi-square test of independence indicated a significant association between the group age and the production of non-plausible responses [$X^2(1) = 11.61$, $p < 0.001$, $V = 0.54$]. According to a Chi-square goodness of fit test, there was a

TABLE 1 | Observed and expected children’s classification according to Siegler rules.

| Groups | Effectives | Siegler rules | | | | |
|----------------|------------|---------------|------|------|-----|-----|
| | | 0 | 1 | 2 | 3 | 4 |
| 4–5-year-olds | Observed | 12 | 6 | 1 | 1 | 0 |
| | Expected | 7.0 | 5.0 | 5.0 | 1.5 | 1.5 |
| 9–10-year-olds | Observed | 2 | 4 | 9 | 2 | 3 |
| | Expected | 7.0 | 5.0 | 5.0 | 1.5 | 1.5 |
| Total | Observed | 14 | 10 | 10 | 3 | 3 |
| | Expected | 14.0 | 10.0 | 10.0 | 3.0 | 3.0 |

significant number of 4–5-year-olds children who produced non-plausible responses ($N = 9$) [$X^2(1) = 4.036$, $p < 0.001$]. A Friedman analysis found no significant difference between the types of torque in non-plausible responses among 4–5-year-olds [$X^2(2) = 1.23$, $p = 0.54$].

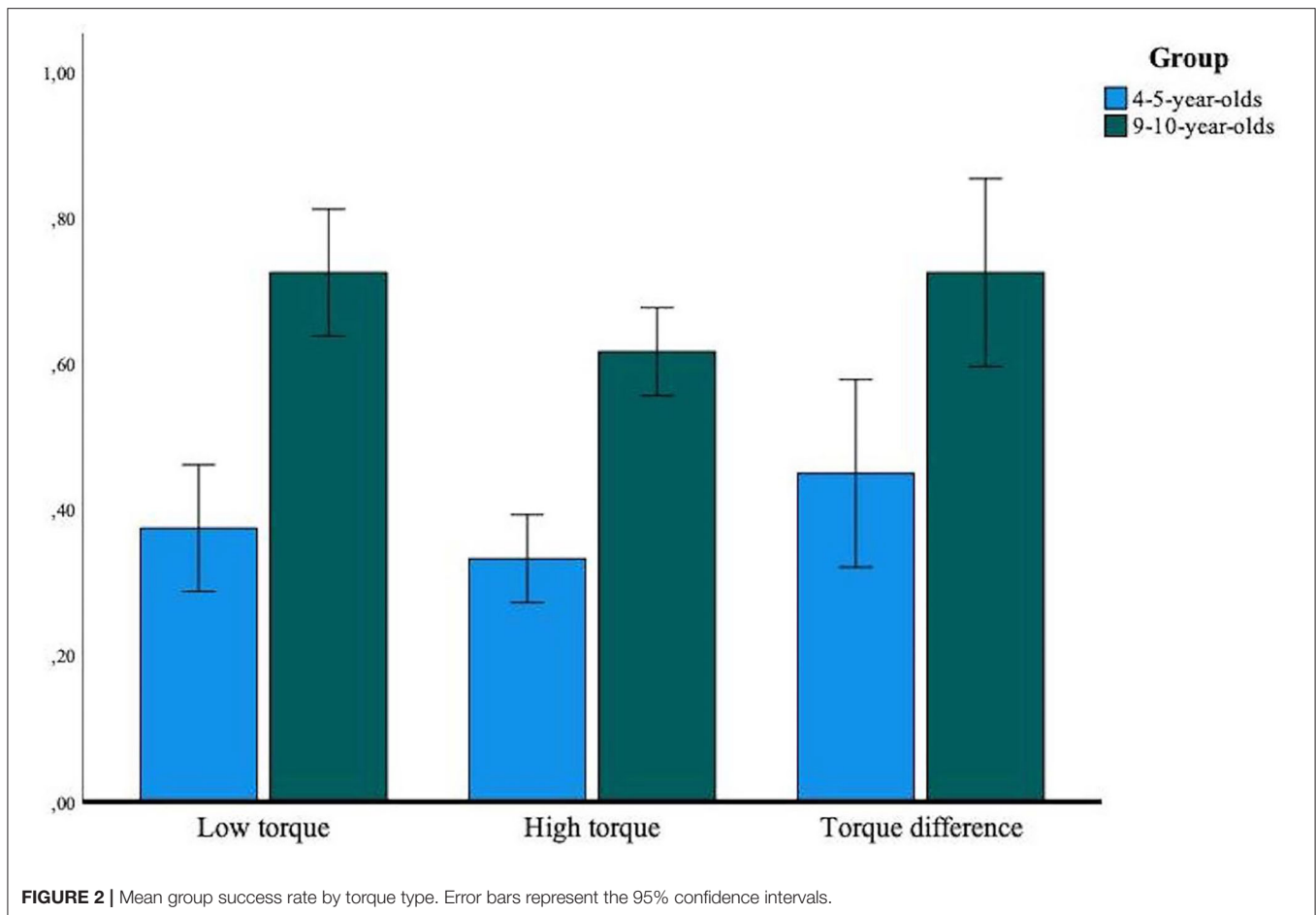
Figure 2 presents the mean success rate by torque type for both age groups. A mixed ANOVA indicated no significant interaction between types of torque and group, [$F_{(2, 76)} = 0.52$, $p = 0.56$, $\eta^2 = 0.01$]. Planned contrasts were used to assess the differences between the types of torque. LT ($M = 0.55$, $SD = 0.19$) success rate was not significantly different than TD ($M = 0.59$, $SD = 0.19$) success rate, [$F_{(1, 38)} = 0.74$, $p = 0.40$, $\eta^2 = 0.02$]. However, LT success rate differed significantly from HT ($M = 0.48$, $SD = 0.19$), [$F_{(1, 38)} = 7.28$, $p < 0.01$, $\eta^2 = 0.16$].

Figure 3 shows the 41 non-plausible responses of the younger children as a function of the position of the bell. The association between non-plausible answers and the position the bell was tested with a Chi-square goodness of fit test. “TotalDown” responses ($N = 36$), differed significantly by bell position [$X^2(1) = 5.44$, $p < 0.05$, $V = 0.39$], but not “TotalUp” ($N = 5$) [$X^2(1) = 0.2$, $p = 0.66$, $V = 0.2$].

DISCUSSION

The present study is consistent with previous findings, as most 4–5-year-olds seemed align with Rules 0 or 1, and 9–10-year-olds with Rules 1 or 2. The impact of weight on the scale is easily understood by children because it is more salient and relevant in their environment and experiences (Inhelder and Piaget, 1958; Siegler, 1976; Ferretti and Butterfield, 1986; Halford et al., 2002; Jansen and van der Maas, 2002). The contribution of distance takes more time to notice and integrate (Leuchter and Naber, 2019). Thus, it is possible to observe Rule 1 until the age of 11 (Siegler and Chen, 1998; Jansen and van der Maas, 2002; Leuchter and Naber, 2019).

Three children in this study were classified with Rule 4. Before age 14, children do not typically understand the torque rule (Siegler and Chen, 1998; Jansen and van der Maas, 2002). Children who resolved the conflict problems could have succeeded by intuition (Messer et al., 2008; Dandurand and Shultz, 2009; Hofman et al., 2015). Children seem to be able to solve the problems without being able to verbalize their reasoning



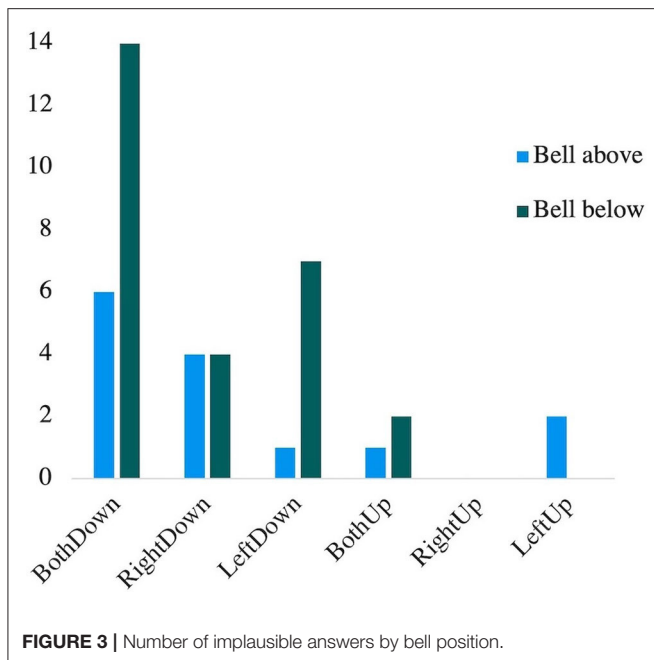
(Messer et al., 2008). Ironically, this is a departure from Piaget's clinical interview method, whereby reasoning is assessed by explanation (Posner and Gertzog, 1982).

The rule classification results we report are important in two respects. First, we deviate from standard approaches used with the balance-scale task by asking children to process each arm independently. Despite this departure, the differences between age groups in our study mirror findings from traditional rule-assessment methods. Arguably, our task measures the same cognitive abilities. Those results suggest that the present method is an adequate version of the balance-scale task. Second, we used an online testing approach to accommodate in-person testing restrictions during the Covid-19 pandemic. Unlike children tested in person, our sample were not able to physically manipulate the balance scale themselves (although they were given three trials to ask the experimenter to manipulate it for them as they saw fit). Rule classification results suggest that online testing can assess similar abilities to what is normally measured in the lab. This is particularly relevant for inclusivity imperatives, whereby online testing can help solve the so-called WEIRD problem in psychology (Jones, 2010). The balance-scale task could be used online to reach typically underrepresented groups. Unfortunately, the minimal demographic information collected in this preliminary study does not allow to assess

inclusion. Further work, with appropriate recruitment strategies, is required to assess inclusiveness targets.

As noted, Siegler's rules approach does not explain some kinds of errors children can produce in the balance-scale task (Sirois et al., 2005; Boom and ter Laak, 2007). In standard studies, children's understanding of the scale and their reasoning is based on the prediction of the movement of the scale using a restricted set of plausible (albeit not necessarily correct) answer choices. Children are presented three possible answers (i.e., right arm goes down, left arm goes down or both arms stay stable). However, in the present study, children had to predict the movement of each arm. They could process one arm independently of the other arm. This methodology can lead to a more detailed understanding of the child's reasoning.

We predicted that younger children who do not understand the unified character of the scale would suggest non-plausible behavior of the scale (an interesting type of error, insofar as characterizing their thinking, that is not allowed in standard task protocols). The older children were not expected to make those errors. The non-plausible responses produced by younger children seem to confirm the hypothesis. There were 9 out of 20 of 4–5-year-olds who provided responses that are implausible. About half of younger children seemed to process each arm independently of the other arm. The misunderstanding of the



unified character of the scale could be caused by the focus of children on their own action. Children aged 3–5 conflate their own actions with those of other objects (Piaget, 1928; Inhelder and Piaget, 1958). In our task, the action of children is divided in two answers, so it could mean that, for them, the balance has also two different actions.

When children misunderstand the unified character of the scale, it appeared associated with a salient feature (i.e., a bell placed above or below the scale). Most non-plausible responses were due to children predicting both arms going down, and were primarily associated with the bell located below the scale. It seems that children understand gravity due to their daily life experiences of the downward pull of weight (Halford et al., 2002). Salient features can lead younger children to focus on a specific aspect of the task (Piaget, 1928, 2002; Amsel et al., 1996). If there was a focus on the transformation (i.e., ring the bell), in relation to their knowledge, this could explain part of the presence of non-plausible responses. It would be interesting to verify this exploratory finding in future studies that use upward force to manipulate a balance-scale, and whether this would be associated with more upward non-plausible responses. At this time, a cautious conclusion is that an incidental salient feature can affect performance on the balance-scale, which could be uniquely useful for a finer-grained analysis of children's understanding.

In previous studies, the saliency caused by the bigger torque on one arm seemed to facilitate the choice between the three possible answers for children (i.e., the torque effect; Ferretti and Butterfield, 1986; Jansen and van der Maas, 2002; Shultz and Takane, 2007; Li et al., 2017). When children's task is to predict the movement of both arms, we expected that the torque effect would not occur for younger children who misunderstand the scale, but that it would be present for older children. Results

suggest that the torque effect fades when children of all ages have to process information from both arms to predict their movement. This effect could be explained by different process for both age group. The encoding ability for younger children is less efficient (Boom and ter Laak, 2007). When multiple stimuli are presented, they do not seem able to encode all information at the same time (Siegler, 1976; Amsel et al., 1996). In standard studies, children can choose which arm to process and ignore information from the other arm. However, when younger children are required to specifically process one arm, they will only take into consideration information from this arm. After, they will process the other arm independently of the first one they processed. Therefore, children could miss the salience of the difference between the two torques because they do not have a global perspective allowing for relative comparisons.

Thus, there is no evidence of a torque effect in the present study. For older children, the success rate for torque difference and low torque were similar. Older children can more easily process information of both arms at the same time (Amsel et al., 1996), but the present task imposed a stepwise reasoning approach. Sometimes, older children gave an answer for one arm and, when they had to give an answer for the second arm, they would change their first answer to ensure a better fit. It happened for most 9–10-year-olds, but it was not documented. It would be useful in future work to include that metric to understand when and how many times children use that strategy. It is possible that a bigger difference between the two arms still facilitated responses, but that low torque is also facilitated because of the methodology. The possibility to take time to process both arms could have increased the success rate of low torque as well.

The high torque trials seem to have a lower success rate than the other types of torque. It could be explained by the perceptual properties of that torque. Both arms are salient in that type of torque. The force applied to the scale on both arms could increase the difficulty of the problems for children of both ages. Children could have made an association between large torque and downward motion. However, with two high-torque arms, it would be relatively difficult to understand the problem, leading to errors.

The interpretation of our findings must be done with caution. The thinking of children seems variable between and within studies according to methodological differences (Halford et al., 2002; Messer et al., 2008; Bullard, 2009; Zon and Xie, 2014). Asking to predict the movement of both arms could explain our results, but it needs to be replicated to assess when children understand the unified nature and behavior of the scale. In a future experiment, it would be useful to add a practice phase after the manipulation phase, as children need time to properly understand a task (Jansen and van der Maas, 2002).

The present study adds information about the nature of children's thinking when the balance-scale task is altered. Perceptual properties of the task do affect children's performance (Halford et al., 2002; Messer et al., 2008). Children base their answers on their intuition, which is substantially about the visual properties of the presented problem (Bullard, 2009). However, those perceptual properties can lead children to errors in their reasoning. Importantly, decades of research with this task may

have overestimated the competence of younger children, as the task demands of standard studies minimize potential errors that have uniquely been revealed in the current study.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Comité d'éthique en recherche avec des êtres humains de l'Université du Québec à Trois-Rivières. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

VF and SS designed the study and performed the statistical analysis and wrote the manuscript. VF prepared the stimuli,

recruited participants, and collected the data. SS wrote the Matlab scripts to process and compile data. All authors contributed to the article and approved the submitted version.

FUNDING

VF was funded by an NSERC Undergraduate Student Research Awards, and is currently funded by a FRQNT masters' scholarship.

ACKNOWLEDGMENTS

We thank all participants' parents because this research could not have happened without their help. We also especially thank the children who took part in the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.702524/full#supplementary-material>

REFERENCES

- Al-Atrash, Y. E., Wishah, A. T., Abul-Omreen, T. H., and Abu-Naser, S. S. (2020). Modeling cognitive development of the balance scale task using ANN. *Int. J. Acad. Information Syst. Res.* 4, 74–81.
- Aldenderfer, M. S., and Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills, CA: Sage. doi: 10.4135/9781412983648
- Amsel, E., Goodman, G., Savoie, D., and Clark, M. (1996). The development of reasoning about causal and noncausal influences on levers. *Child Dev.* 67, 1624–1646. doi: 10.2307/1131722
- Boom, J., and ter Laak, J. (2007). Classes in the balance: Latent class analysis and the balance scale task. *Dev. Rev.* 27, 127–149. doi: 10.1016/j.dr.2006.06.001
- Bullard, D. P. (2009). *The Impact of Context Manipulation on Knowledge Development in a Balancing Task*. University of Cincinnati.
- Dandurand, F., and Shultz, T. (2009). *Modeling Acquisition of a Torque Rule on the Balance-Scale Task*. 31st Annual Meeting of the Cognitive Science Society. Amsterdam. Retrieved from: <https://escholarship.org/uc/item/29x6r5zn>
- Ferretti, R. P., and Butterfield, E. C. (1986). Are children's rule-assessment classifications invariant across instances of problem types? *Child Dev.* 57, 1419–1428. doi: 10.2307/1130420
- Ferretti, R. P., and Butterfield, E. C. (1992). Intelligence-related differences in the learning, maintenance, and transfer of problem-solving strategies. *Intelligence* 16, 207–223. doi: 10.1016/0160-2896(92)90005-C
- Halford, G. S., Andrews, G., Dalton, C., Boag, C., and Zielinski, T. (2002). Young children's performance on the balance scale: the influence of relational complexity. *J. Exp. Child Psychol.* 81, 417–445. doi: 10.1006/jecp.2002.2665
- Hofman, A. D., Visser, I., Jansen, B. R., and van der Maas, H. L. (2015). The balance-scale task revisited: a comparison of statistical models for rule-based and information-integration theories of proportional reasoning. *PLoS ONE* 10:136449. doi: 10.1371/journal.pone.0136449
- Inhelder, B., and Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence: An Essay on the Construction of Formal Operational Structures*. New York, NY: Basic Books. doi: 10.1037/10034-000
- Jansen, B. R., and van der Maas, H. L. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Dev. Rev.* 17, 321–357. doi: 10.1006/drev.1997.0437
- Jansen, B. R., and van der Maas, H. L. (2002). The development of children's rule use on the balance scale task. *J. Exp. Child Psychol.* 81, 383–416. doi: 10.1006/jecp.2002.2664
- Jones, D. (2010). A weird view of human nature skews psychologists' studies. *Science* 328:1627. doi: 10.1126/science.328.5986.1627
- Leuchter, M., and Naber, B. (2019). Studying children's knowledge base of one-sided levers as force amplifiers. *J. Res. Sci. Teach.* 56, 91–112. doi: 10.1002/tea.21470
- Li, F., Xie, L., Yang, X., and Cao, B. (2017). The effect of feedback and operational experience on children's rule learning. *Front. Psychol.* 8 :534. doi: 10.3389/fpsyg.2017.00534
- Messer, D. J., Pine, K. J., and Butler, C. (2008). Children's behaviour and cognitions across different balance tasks. *Learn. Instruct.* 18, 42–53. doi: 10.1016/j.learninstruc.2006.09.008
- Piaget, J. (1928). La causalité chez l'enfant. *Br. J. Psychol.* 18:276. doi: 10.1111/j.2044-8295.1928.tb00466.x
- Piaget, J. (2002). "The epigenetic system and the development of cognitive functions," in *Brain Development and Cognition: A Reader (Second Edition)* (Oxford: Blackwell), 29–35. doi: 10.1002/9780470753507.ch3
- Posner, G. J., and Gertzog, W. A. (1982). The clinical interview and the measurement of conceptual change. *Sci. Educ.* 66, 195–209. doi: 10.1002/sce.3730660206
- Richardson, F. M., Baughman, F. D., Forrester, N. A., and Thomas, M. S. (2006). "Computational modeling of variability in the balance scale task," in *Proceedings of the 7th International Conference of Cognitive Modeling* (Trieste).
- Shultz, T. R. (2017). "Constructive artificial neural-network models for cognitive development," *New Perspectives on Human Development*, eds N. Budwig, E. Turiel, and P. D. Zelazo (Cambridge University Press), 15–25. doi: 10.1017/CBO9781316282755.003
- Shultz, T. R., Mareschal, D., and Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learn.* 16, 57–86. doi: 10.1007/BF00993174
- Shultz, T. R., and Takane, Y. (2007). Rule following and rule use in the balance-scale task. *Cognition* 103, 460–472. doi: 10.1016/j.cognition.2006.12.004
- Siegler, R. S. (1976). Three aspects of cognitive development. *Cogn. Psychol.* 8, 481–520. doi: 10.1016/0010-0285(76)90016-5

- Siegler, R. S. (2016). Continuity and change in the field of cognitive development and in the perspectives of one cognitive developmentalist. *Child Dev. Perspect.* 10, 128–133. doi: 10.1111/cdep.12173
- Siegler, R. S., and Chen, Z. (1998). Developmental differences in rule learning: a microgenetic analysis. *Cogn. Psychol.* 36, 273–310. doi: 10.1006/cogp.1998.0686
- Siegler, R. S., and Chen, Z. (2002). Development of rules and strategies: balancing the old and the new. *J. Exp. Child Psychol.* 81, 446–457. doi: 10.1006/jecp.2002.2666
- Siegler, R. S., and Richards, D. D. (1979). Development of time, speed, and distance concepts. *Dev. Psychol.* 15 :288. doi: 10.1037/0012-1649.15.3.288
- Sirois, S., Markovits, H., and Pomerleau-Laroche, M. E. (2005). "Manipulating children's errors on the balance-scale task," in *XIIth European Conference on Developmental Psychology*. La Laguna, Tenerife.
- Zimmerman, C. L. (1999). *A Network Interpretation Approach to the Balance Scale Task*. Edmonton, AB: University of Alberta.
- Zon, T. W., and Xie, B. F. (2014). *Using Artificial Neural Networks to Model Siegler's Balancing Task*. Swarthmore, PA: Swarthmore College. Available online at: <https://www.cs.swarthmore.edu/~meeden/cs81/s14/papers/BenTyler.pdf>

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Filion and Sirois. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Quantifying Everyday Ecologies: Principles for Manual Annotation of Many Hours of Infants' Lives

Jennifer K. Mendoza and Caitlin M. Fausey*

Department of Psychology, University of Oregon, Eugene, OR, United States

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Margaret Moulson,
Ryerson University, Canada
Melanie Soderstrom,
University of Manitoba, Canada

*Correspondence:

Caitlin M. Fausey
fausey@uoregon.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 16 May 2021

Accepted: 20 July 2021

Published: 06 September 2021

Citation:

Mendoza JK and Fausey CM (2021)
Quantifying Everyday Ecologies:
Principles for Manual Annotation of
Many Hours of Infants' Lives.
Front. Psychol. 12:710636.
doi: 10.3389/fpsyg.2021.710636

Everyday experiences are the experiences available to shape developmental change. Remarkable advances in devices used to record infants' and toddlers' everyday experiences, as well as in repositories to aggregate and share such recordings across teams of theorists, have yielded a potential gold mine of insights to spur next-generation theories of experience-dependent change. Making full use of these advances, however, currently requires manual annotation. Manually annotating many hours of everyday life is a dedicated pursuit requiring significant time and resources, and in many domains is an endeavor currently lacking foundational facts to guide potentially consequential implementation decisions. These realities make manual annotation a frequent barrier to discoveries, as theorists instead opt for narrower scoped activities. Here, we provide theorists with a framework for manually annotating many hours of everyday life designed to reduce both theoretical and practical overwhelm. We share insights based on our team's recent adventures in the previously uncharted territory of everyday music. We identify principles, and share implementation examples and tools, to help theorists achieve scalable solutions to challenges that are especially fierce when annotating extended timescales. These principles for quantifying everyday ecologies will help theorists collectively maximize return on investment in databases of everyday recordings and will enable a broad community of scholars—across institutions, skillsets, experiences, and working environments—to make discoveries about the experiences upon which development may depend.

Keywords: annotation, input, music, infancy, LENA

INTRODUCTION

Experience-dependent changes in neural circuitry and behavior are central to development (Hensch, 2005; Hannon and Trainor, 2007; Scott et al., 2007; Aslin, 2017). Complete theories of development must therefore model the experiences that drive change. In human infancy, detailed models of real-world early experiences have traditionally been hard to achieve because of challenges in recording everyday sensory histories. Recent technological advances that permit many hours of recording have minimized this barrier of experience sampling *per se* (de Barbaro, 2019). One insight from recent efforts using these technologies is that when relevant sensory histories naturally unfold over extended timescales, shorter samples miss pervasive properties of infants' everyday ecologies. For example, short language samples miss typical rhythms of interleaving speech and silence (Tamis-LeMonda et al., 2017; Cristia et al., 2021) and short musical samples

(Mendoza and Fausey, 2021a) fail to capture opportunities for repetition and variability as instances arise non-uniformly over time (Smith et al., 2018). An emerging priority for theories of development is therefore to model very large amounts of everyday experience. Though recording such quantities is now possible, automatically detecting relevant units within the complex and varied sensory streams of everyday life is not (Adolph, 2020; de Barbaro and Fausey, in press). Developmental theorists must therefore tackle the challenge of manually annotating many hours of everyday life.

Manually annotating many hours of everyday life is so daunting that most researchers who have recorded such data avoid it. The status quo is to declare longform manual annotation “impractical,” “untenable,” “not realistic,” “challenging,” and “unwieldy” (Roy et al., 2015; Casillas et al., 2017; Tamis-LeMonda et al., 2018; Casillas and Cristia, 2019; Räsänen et al., 2019). Despite developmental theorists’ considerable expertise in annotating behavior (Bakeman and Gottman, 1997; Adolph, 2020), scaling from researcher-constrained short activities to everyday ecologies is not straightforward. One challenge is that everyday sights and sounds are not just “more” data, but also “different” data. Theorists must update operationalizations of annotation targets based on new and variable instantiations arising in everyday sensory streams. Another challenge is a lower signal-to-noise ratio in everyday contexts compared to researcher-constrained contexts because of multiple overlapping sources generating the sensory streams. Audio data, in particular, are often literally “noisier” (Xu et al., 2009). Reaching conventional thresholds for reliably identifying annotation targets is therefore a Sisyphean task that often demands updated rationale. Finally, because successful annotation requires manyfold the duration of the annotated recording (MacWhinney, 2000), manually annotating many hours of everyday life requires very large investments of time, personnel, and dedicated resources including sustained funding (Casillas and Cristia, 2019; VanDam and De Palma, 2019). Theorists must achieve remarkable “operations manager” prowess in their laboratories. This suite of challenges is fierce but it need not thwart research progress. Here, we articulate principles for manually annotating many hours of everyday life that minimize challenges and maximize opportunities for new discoveries about infants’ everyday ecologies.

The potential for new discoveries about infants’ everyday ecologies is perhaps higher than ever, given repositories of everyday experiences like Databrary (<https://nyu.databrary.org/>; Gilmore et al., 2018) and HomeBank (<https://homebank.talkbank.org/>; VanDam et al., 2016). Each of these repositories already contains many hours of recordings captured from infants’ everyday lives that are available for theorists to annotate. Regularities in everyday audio, including multiple levels of vocalization, language, and music, as well patterns in multi-modal video including emotional expressions, contingencies and motor dynamics among social partners, and nameable object and actions, are hypothesized to shape developmental change. Thus, quantifying these everyday regularities will inform developmental theory including computational models that currently lack everyday parameters.

As a scientific process, manual annotation of many hours of everyday life is also well-suited to priorities like expanding our scientific workforce by including people, expertises, and institutions who have traditionally faced systemic barriers to participation in discovery. For example, though not every investigator may always have resources to innovate technology or to collect massive samples of new data, a very large number of scientists and their teams can conduct manual annotation of already existing everyday data. Further, the opportunity to aggregate across diverse samples of everyday data—each individual corpus in Datavyu and HomeBank is necessarily limited by space, time, and community—demands theorists’ engagement in order to determine the extent to which findings vary across cultural contexts (Nielsen et al., 2017; Hruschka et al., 2018; Cychosz et al., 2020a; Soderstrom et al., 2021). Finally, as we experience disruptions like the COVID-19 pandemic and other barriers to traditional laboratory business-as-usual, manual annotation of many hours of everyday recordings is a scientific endeavor that is both feasible and likely to yield theory-relevant insights. Manual annotation is a classic bottleneck in maximizing returns on scientific investments, especially when initial study design and data collection generate very large datasets of continuous recordings of infants’ everyday ecologies. Manually annotating these everyday data will yield theoretical insights as well as create goldstandard training and evaluation sets en route to eventual automated annotation (Bambach et al., 2016; Ossmy et al., 2018; Räsänen et al., 2019). Given our own team’s recent adventures, we share here critical reflections on practices for manual annotation of many hours of everyday lives likely to advance developmental theory.

We share seven principles and materials to support their implementation (osf.io/eb9pw, henceforth “OSF”; Mendoza and Fausey, 2019) based on our recent discoveries about everyday music in infancy (Fausey and Mendoza, 2018a,b; Mendoza and Fausey, 2018, 2021a,b). Briefly, we audio recorded 35 full days-in-the-lives-of-infants and then identified the musical features, voices, and tunes available over the course of each day. Because music cannot yet be automatically detected in recordings of everyday life (Mehr et al., 2019), we pre-processed and then double-annotated roughly 270 hours of everyday audio. Among other findings, we discovered that infants encounter roughly 1 h of music per day, a quarter of which is live and vocal, with some musical tunes and voices preferentially available. So that other theorists can build on these discoveries, and so that scholars across domains can tackle manually annotating many hours of everyday life, here we present a framework for guiding the many decisions in a manual annotation workflow.

Music is an illustrative domain because there is very little extant evidence to inform decisions about manual annotation. The early days of a discovery process—the situation in which most theorists find themselves when first scaling to quantify everyday ecologies—present distinct challenges for justifying analytic decisions. All-day is an illustrative timescale because days constrain activities and their accompanying sensory details (Hofferth and Sandberg, 2001; Galland et al., 2012; Roy et al., 2015; Montag et al., 2018), 16-h audio days are feasible to record (Ford et al., 2008; Ganek and Eriks-Brophy, 2018),

and yet few discoveries about everyday experiences have harnessed this extended timescale (though see Soderstrom and Wittebolle, 2013; Weisleder and Fernald, 2013; Roy et al., 2015 for related approaches). We add to a growing set of resources designed to support manual annotation, like the CHAT manual for transcribing language (talkbank.org; MacWhinney, 2000), Datavyu and ELAN for annotating and analyzing audio and/or video data (Wittenburg et al., 2006; Datavyu Team, 2014) and the DARCLE and ACLEW Annotation Schemes (Casillas et al., 2017; Soderstrom et al., 2021) for annotating speech in prioritized subsets of daylong audio recordings. We emphasize the conceptual and implementation needs associated with manually annotating many hours of everyday life.

We share a set of principle-implementation pairs (**Figure 1**). We prioritize theorists' agency and so share a framework to structure decision-making rather than prescribing step-by-step instructions *per se*. Principles 1–3 address fundamental decisions about what to annotate in many hours of everyday life. Principles 4–6 address how to achieve reliable annotations at scale. Principle 7 addresses infrastructure for successful annotation. Each section of this paper presents one principle-implementation pair, first articulating the theoretical issues at stake and then describing implementation procedures. We share associated files like coding manuals and scripts on OSF to demystify the process and facilitate future efforts in the ambitious endeavor of making discoveries about infants' everyday ecologies.

PRINCIPLE 1: INCLUSIVE AND HIERARCHICAL INSTANTIATIONS OF CONSTRUCTS IN MANY HOURS OF EVERYDAY LIFE

The core goal of manual annotation is to identify annotation targets within the stream of sensory experiences captured by a recording device. Because sensory histories are not uniform (e.g., Jayaraman et al., 2015; Roy et al., 2015; Tamis-LeMonda et al., 2017; Clerkin et al., 2018; Mendoza and Fausey, 2021a), annotation targets will sometimes be present and other times be absent throughout the recorded stream. Discovering structure in this everyday ecology requires identifying when instances of the “same” thing happen again even when separated in time, context, and with only partially overlapping instantiations. For example, over the course of a day, an infant might encounter a parent singing the first phrases of “Twinkle, Twinkle Little Star” at 8 a.m., a cartoon character singing “Wheels on the Bus” at 8:30 a.m., and then the same parent singing the entire “Twinkle, Twinkle Little Star” at 6 p.m. All of these example instances are music, all are vocal, some are live, some are recorded, and the voice and tune identities partially overlap across instances. Each kind of repetition and variation may be relevant for building musical skills like detecting multiple levels of musical structure, recognizing melodies, and generalizing musical meter (e.g., Hannon and Trehub, 2005; Margulis, 2014; Creel, 2019). Other potentially musical sounds may also arise in the infant's day, including some whistling, speech sound effects like “beep beep,” and clapping. Segmenting the stream of everyday experience so

that repetitions and variations are discoverable requires detailed operationalizations of annotation targets.

The challenge when manually annotating many hours of everyday life is to achieve operationalizations that faithfully capture the everyday phenomena. Theorists often have little direct evidence from everyday life to guide decisions (one notable exception is language, in which “words” have long been recognized as important and transcribable units; MacWhinney, 2000). Extant evidence from researcher-constrained experiments often suggests relevant starting points. For example, musical sounds have been defined as “humanly produced, non-random sequences of tones or tone combinations that are non-referential” and vocal, instrumental, live, and recorded music have been instantiated in laboratory tasks (Trehub and Schellenberg, 1995). Infants also track repetition and variation across musical tunes and voices in these tasks (for review see Mendoza and Fausey, 2021a). Researcher-constrained instantiations of a construct are only starting points for operationalizing annotation targets because everyday sensory streams include a much wider range of activities, behaviors, and generating sources than laboratory contexts (Young and Gillen, 2007; Lamont, 2008; de Barbaro and Fausey, in press). For example, is vocal play from siblings (affectionately referred to as “scream singing” by our team) music? How about microwave beeping? Do humming and clapping deserve the same status as singing a complete rendition of “Happy Birthday”? Does half a rendition count? Many real-life instantiations of a construct have never been measured or manipulated in the laboratory (e.g., Lee et al., 2018). In the face of unknowns and potentially highly variable instantiations of to-be-annotated constructs over the course of many hours of everyday life, it is productive to operationalize a construct inclusively. Theorists can later quantify the prevalence and structure of specific subsets in future annotation efforts.

Multiple passes of manual annotation allow theorists to quantify a planned hierarchy of construct instantiations. For example, after inclusively annotating “music,” annotators can later identify features, voices, and tunes within the music, and then achieve even finer-grained transcriptions of pitches and rhythms. When very little about everyday ecologies is known, systematic manual annotation from more-general to less-general permits initial insights based on aggregating operationalizations across prior researcher-constrained investigations. This inclusive first pass thus identifies everyday structure that is broadly relevant to cumulative science. Insights about more specific instantiations (e.g., “live vocal complete renditions of tunes in C major”) may be difficult to initially discover given the sparsity of individual types in everyday ecologies (Zipf, 1936, 1949). If present, these instantiations can be quantified in subsequent annotation passes of inclusively annotated constructs.

There is no single instantiation of a construct across many hours of everyday life. There may not even be predictable deviations from a prototypical instantiation derived from prior evidence based on researcher-constrained tasks. The opportunity for discovery in everyday ecologies is real and so specific instantiations of constructs cannot always be entirely pre-planned. In order to operationalize inclusive and hierarchical instantiations of annotation

| |
|--|
| Principle 1: Inclusive and hierarchical instantiations of constructs in many hours of everyday life OSF component: Coding Manuals Subcomponents: Identifying music bouts in daylong recordings Tagging features, voices, & tunes in music bouts |
| Principle 2: Theory-informed scheme for sampling many hours of everyday life OSF component: Data Processing Subcomponent: Recording pre-processing procedures and code |
| Principle 3: Finest-grained defensible timescale of observed phenomena in many hours of everyday life OSF component: Data Processing Subcomponent: Generating music, features, voices, and tunes data |
| Principle 4: Transparent rationale for assessing reliability of annotations in many hours of everyday life OSF component: Data Processing Subcomponent: Inter-rater reliability procedures and code |
| Principle 5: Active and representative training protocol for annotating many hours of everyday life OSF components: Coder Training Coding Manuals Subcomponents: Training to code music bouts Training to code features, voices, & tunes Identifying music bouts in daylong recordings Tagging features, voices, & tunes in music bouts |
| Principle 6: Procedural priority on selective attention when annotating many hours of everyday life OSF components: Coder Training Coding Manuals Subcomponents: Training to code music bouts Training to code features, voices, & tunes Identifying music bouts in daylong recordings Tagging features, voices, & tunes in music bouts |
| Principle 7: Inclusive physical, social, and professional communities for annotators of many hours of everyday life OSF component: Coder Training Subcomponent: Welcome to the Learning Lab |

FIGURE 1 | Principles for manual annotation of many hours of infants' everyday lives. Each principle guides decisions in a research program aiming to quantify everyday ecologies. Materials that support implementing these principles, instantiated in a line of research about infants' everyday music, are shared on OSF (<https://osf.io/eb9pw/>; Mendoza and Fausey, 2019). For an accessible version, please go to <https://osf.io/vd8t5>.

targets in many hours of everyday life, theorists must therefore (1) conduct iterative pilot annotation, (2) create comprehensive definitions, and (3) find unambiguous examples for training.

Implementation

We identified the onset and offset of individual instances of music (“music bouts”) in many hours of everyday life. Music bouts were operationalized inclusively. We then identified whether each music bout included live, recorded, vocal, and/or instrumental music. We also identified individual tune(s) and voice(s) in each music bout. After identifying the durations of music bouts, subsequent features were annotated as present or absent per bout given that durations would arise straightforwardly for those bouts containing a single feature (see Mendoza and Fausey, 2021a; we repeat some methodological rationale throughout). **Figure 1** shows the OSF components that supported this annotation as we took the following steps in our workflow. One illustrative file is “3_FeaturesVoicesTunes_CodingManual.pptx” (<https://osf.io/dtfnv/>) which is the coding manual for identifying musical features, voices, and tunes in many hours of everyday life.

Iterative Pilot Annotation

Because many hours of everyday life present opportunities to discover previously unobserved instantiations of constructs, iterative “annotate-discuss-update” loops should inform eventual operationalizations that will guide the annotation of a planned sample of recordings. Theorists can avoid endless iterations by selecting pilot recordings from distinct family contexts and by engaging pilot annotators who have varying levels of expertise about children and the target domain. To develop our operationalizations for annotating everyday music, we completed several iterations of annotating recordings collected for this purpose. We solicited feedback from annotators about sounds that were easy or hard to identify as “music” in everyday ecologies. This process revealed the range of pitched and rhythmic sounds in a typical day of infants' lives. We discovered it was necessary to define not only the range of sounds that should be considered “music,” but also the range of sounds that should not be considered “music.” Our operationalizations were consistent with, and yet more specific and varied than, definitions in extant literature. We also annotated pilot recordings en route to everyday-appropriate operationalizations of musical features, voices, and tunes. Pilot annotators during this phase noted partial instantiations of standard tunes. For example, the pitch patterns

in the first phrase (“Twinkle, twinkle little star”) and in the second phrase (“How I wonder what you are”) of “Twinkle, Twinkle Little Star” are different. If these phrases occurred in distinct musical bouts, would each bout receive the same tune annotation (“Twinkle, Twinkle Little Star”)? Reasonable theorists could arrive at different conclusions; we therefore emphasize making these decisions transparent by sharing detailed coding manuals and data so that future efforts can assess the extent to which such decisions matter for pattern discovery. Importantly, iterative pilot annotation revealed everyday instantiations of music and its features, voices, and tunes that were essential to address in annotation manuals.

Comprehensive Definitions

We arrived at a three-part definition of music that specified (1) the range of sounds that should be coded as music, (2) the range of sounds that should not be coded as music, and (3) the start and end of music bouts. We also clearly defined musical features (i.e., live, recorded, vocal, and/or instrumental music), voices, and tunes (**Supplementary Table 2**, OSF, <https://osf.io/htx57/>). In each definition, we highlighted the range of possible types that annotators might encounter. For example, our definition of vocal music lists several possible kinds of voices (e.g., adult, non-focal child, recorded character) and also the different types of vocal music (i.e., singing, humming, whistling, vocal play). We intentionally created definitions that emphasized the variability of instances that should be annotated.

Unambiguous Audio Examples for Training

In our two coding manuals, we combined the comprehensive definitions of music and musical properties with clear audio examples, extracted from pilot recordings, to illustrate each to-be-annotated phenomenon. We used audio examples that unambiguously depicted our phenomena. These prototypical anchors helped annotators decide what to do when they encountered an everyday sound in daylong audio recordings that was hard to annotate. For example, infants’ older siblings commonly produced a very wide range of vocalizations, only some of which should be considered “music” under our annotation scheme. When an annotator encountered a sibling vocalization that they were not sure about, they could listen to the full set of audio examples that should be coded as “music” and the full set of audio examples that should not be coded as “music” and then decide to which set the specific sibling vocalization was most similar.

PRINCIPLE 2: THEORY-INFORMED SCHEME FOR SAMPLING MANY HOURS OF EVERYDAY LIFE

Theorists sample from everyday life when they record it and when they annotate it. Theorists must therefore choose how much and when to sample. The goal is to sample in such a way that allows theorists to make discoveries about everyday ecologies that both respect things we already know and move us in some way beyond what we currently know. Currently, we know very little about the prevalence and rhythm of various

sensory events in everyday life. This simultaneously licenses an exploratory mindset, in which some insights are better than no insights so that an empirical foundation can take shape over time, as well as strategic considerations of what could make for the highest yield insights upon recording or annotating infants’ everyday ecologies. Central to these considerations is the multi-scale nature of time. Though sampling decisions are often posed as decisions about a single timescale—“Should I record 1 h, 1 day, 1 week, 1 month, or 1 year? Should I annotate all minutes or just some minutes of each recording?”—the reality is that briefer timescales are always nested within more extended timescales and attention, memory, and learning mechanisms operate over multiple timescales as infants build knowledge (Thelen and Smith, 1994). Thus, because theories of experience-dependent change require evidence from multiple timescales of everyday experiences, it is productive to consider the extent to which any sampling decision yields a “multi-scale dividend” by potentiating insights at multiple theory-relevant timescales.

Theorists designing sampling plans face a classic conundrum in that the prevalence of their target behavior constrains optimal sampling yet prevalence itself is often unknown (Collins and Graham, 2002; Adolph et al., 2008). Importantly, prevalence at one timescale does not straightforwardly predict prevalence at other timescales. For example, individual instances of many behaviors like bouts of walking, attention to objects, and music often last on the order of seconds (Adolph et al., 2012; Suarez-Rivera et al., 2019; Mendoza and Fausey, 2021a,b) and these brief instances do not arise at a steady rate across an hour or across a day. One clear illustration of this is a pattern of speech interleaved with extended periods of silence in samples of everyday audio (Tamis-LeMonda et al., 2017; Cristia et al., 2021). That is, speech rate was not constant but rather rose and fell over the course of the extended recording. Non-uniform temporal rhythms make the endeavor of identifying a rate in a shorter sample and then linearly extrapolating to estimate its rate over longer timescales potentially suspect. Relatedly, interpolating between coarsely timed samples yields trajectories that are meaningfully distorted compared to denser sampling (Adolph et al., 2008). Thus, sampling briefly (e.g., 1 min total) or sparsely (e.g., 1 min per hour) is not likely to yield a multi-scale dividend (e.g., discoveries about secondly, minutely, and hourly prevalence). In contrast, densely annotating many hours of everyday life makes it possible to discover structure at the finest-grained annotated timescale as well as every coarser timescale up to total sampled duration of extended recordings “for free” (see also Principle 3).

Of course, few practicing developmental theorists would consider densely annotating many hours of everyday life “free.” The massive investment of person-hours required for manual annotation costs time and money; the following considerations can inform sampling decisions when balancing feasibility with ambitions of a multi-scale dividend.

Although we have a lot to learn about the prevalence of everyday behaviors, existing evidence often provides some anchors. For example, time-use and retrospective surveys completed by caregivers of young children suggest broad contours of everyday rates (Hofferth and Sandberg, 2001; PSID-CDS User Guide, 2002), such as daily occurrence for music

(Custodero and Johnson-Green, 2003) and weekly rhythms for some aspects of affect and sleep (Larsen and Kasimatis, 1990; Szymczak et al., 1993). Ongoing research using complementary methods like dense Ecological Momentary Assessment in which caregivers report in-the-moment snapshots of their infants' experiences over days, weeks, and months (Franchak, 2019; Kadooka et al., 2021) will also reveal the prevalence of many motor, visual, and language behaviors at extended timescales of everyday life. One's sampling scheme can respect available evidence by sampling at least as densely as known rates, and go beyond extant knowledge by combining any of several denser and/or more extended samples. When everyday prevalence is unknown or coarsely estimated, many timescales (not just the most costly) would yield multi-scale dividends to advance theories of experience-dependent change.

Recent and ongoing efforts are also teaching us about the consequences of various sampling schemes for estimating rates of everyday behaviors within extended recordings. For example, random sampling approximates rates of continuously annotated behaviors when those behaviors are medium or high base rate (Micheletti et al., 2020). Estimates for each of two available languages in everyday speech, as well as rates of adult- and child-directed speech, stabilize upon cumulating roughly 90 min of 30-s segments randomly sampled from a day (Cychosz et al., 2020b). Behaviors with low everyday base rates present the biggest challenge for sampling; erring on the side of continuous annotation is wise for initial efforts that can then inform future sampling schemes. Another productive option is to combine multiple sampling choices such as randomly selected segments together with segments of peak theory-relevant activity (Casillas et al., 2020).

A related consideration is to identify which portions of everyday experience you must quantify in order to best address your primary research question. If sensory input during waking hours is the theory-relevant experience, then samples can be scheduled according to known waking hours per day for infants of various ages (Galland et al., 2012) instead of sampling full 24-h cycles or including mid-day naps. Portions of extended recordings like episodes dense with adult speech and therefore potential social interactions (Ramírez-Esparza et al., 2014; Romeo et al., 2018), and episodes like mealtimes that provide learning opportunities for many early learned object names (Clerkin et al., 2018), are highly relevant for many theories of experience-dependent learning. Here, theorists need only be mindful of extrapolation and interpolation missteps when using such samples to inform estimates of cumulative experience (see also Montag et al., 2018). Thus, theorists can make principled decisions about sampling schemes most likely to achieve a combination of advancing theory, avoiding estimation traps, and feasibility.

Some research questions demand large quantities of everyday data that may vexingly resist attempted downsizing via shorter and/or sparser sampling. Two examples include aiming to understand temporally extended schedules *per se*, due to their hypothesized relevance for learning mechanisms related to spacing and/or sleep consolidation (e.g., Rovee-Collier, 1995; Gómez and Edgin, 2015; Vlach, 2019) and estimating extended

cumulative experiences like functions relating word tokens and types in order to understand everyday lexical diversity (Montag et al., 2018). Other research questions require capturing many instances of everyday sights and sounds (e.g., objects, words, musical tunes, speaker/singer identities, etc.) in order to understand opportunities for learners to encounter repeated and varying instances within and across categories. Accumulating multiple instances often requires extended sampling because everyday behaviors may preferentially occur in particular activities (e.g., breakfast) and on particular days of the week (e.g., only on Saturday). One dramatic illustration of this is the discovery of a total of 313 instances of the word "breakfast" in 15 months of continuously transcribed adult speech (Roy et al., 2015) which works out to fewer than one instance per day. Multiple and varied instances of other behaviors can be quantified by dense annotation within daily or hourly samples (e.g., Clerkin et al., 2018; Tamis-LeMonda et al., 2018; Mendoza and Fausey, 2021a). Altogether, if the necessary volume of everyday instances is unlikely to occur all at once or if there is not yet enough known about a phenomenon to predict when it will occur at a high volume, then theorists may need to sample extended and densely in order to discover theory-relevant distributions of experience.

Sampling is fundamentally a multi-scale matter; 10 min of a morning at home cannot represent the entire life from which it was sampled, and it might not meaningfully represent the month, day, or even hour from which it was sampled. The implications of any particular everyday sample for theories of experience-dependent change will become clearer as theorists identify patterns of relative stability and change at multiple timescales of everyday experience. Measures like coefficient of variation (Anderson and Fausey, 2019), multi-scale coefficient of variation (Abney et al., 2017a), and intra-class correlations (Bolger and Laurenceau, 2013; d'Apice et al., 2019; Mikhelson et al., 2021) all yield insights about these dynamics. Recent investigations have quantified such everyday dynamics from hour-to-hour, day-to-day, and month-to-month (e.g., Fausey et al., 2016; Anderson and Fausey, 2019; d'Apice et al., 2019) and additional insights will increasingly be possible thanks to shared corpora of many hours of infants' everyday lives.

One way to cumulate insights across timescales is to design sampling schemes with extant evidence in mind, taking care to articulate how one's scheme will yield discoveries at briefer and/or more extended timescales than currently known. Another way to potentiate multi-scale insights is to densely (not sparsely) annotate many (not few) hours of everyday life so that analyses can quantify multiple coarser-than-annotated rhythms. Determining timescales of relative stability (e.g., 10 min at the beginning and end of a day may be interchangeable) and relative change (e.g., 1 month sampled at the beginning and end of a year may not be interchangeable) for everyday phenomena will also enable greater precision in relating trajectories of experiences and developmental change. Though dense sampling of extended timescales is not a unique path to insights about infants' everyday ecologies, the multi-scale dividend for such efforts is very high and thus worth pursuing for cumulative science.

Implementation

We extended knowledge about everyday music in infancy by creating a scheme for when and how much audio to record from everyday life, as well as when and how much audio to annotate from within captured recordings. Multiple resources relevant for implementing this principle can be found in the OSF components specified in **Figure 1**. One illustrative file is “Silence_Praat_Loop” (<https://osf.io/egmbh/>), a Praat script that accomplishes a pre-processing step designed to address situations in which families occasionally turned off the LENA™ digital language processor (DLP). In order to ensure that time in each .wav file represents time elapsed during the recorded day, this script inserts silence into .wav files for the duration of any periods when the DLP had been turned off. For example, if a family recorded from 8 a.m. until 8 p.m. and they turned off the DLP from 9 a.m. to 10 a.m., then the resulting .wav file would be 11 h instead of 12 h duration. Inserting 60 min of silence starting 1 h from the beginning of the .wav file preserves continuous time in to-be-annotated files.

Decide When and How Much to Record

We made a theory-informed decision to sample three full days per family distributed within 1 week. Prior research suggested that caregivers would sing and/or play music daily with their infants (Custodero and Johnson-Green, 2003), but there was not yet enough known to predict when music would occur during the day. We therefore sampled densely by instructing families to record the maximum 16-h duration of the LENA™ digital language processor. We sampled multiple days per family in order to potentiate insights about the stability and variability of everyday music across multiple timescales. Three days was based on what would be feasible for families to complete (Gilkerson et al., 2015; Canault et al., 2016).

Decide When and How Much to Annotate

We made a theory-informed decision to densely sample the many hours of recorded life. One of our research aims was to discover the total duration of music per day in infants' lives. Because the prevalence and timing of music bouts within a day were unknown, we annotated continuously in order to detect each bout. This approach yielded 42 h of everyday music from within 467 h of everyday sounds. We also aimed to quantify the repetition and variation of features, voices, and tunes within everyday music. Prior research suggested that unique instantiations of music might occur sparsely during infants' days (Costa-Giomi and Benetti, 2017) and so continuous annotation was most likely to identify the full range of the day's musical features, voices, and tunes.

Decide What Not to Annotate

We made a theory-informed decision not to annotate long stretches of silence or very low-level sounds since these portions of the recordings were unlikely to contain our phenomenon of interest. Our approach for identifying and excluding these portions of the recording is generalizable to studying other auditory phenomena and consistent with pre-processing steps used in prior research (Weisleder and Fernald, 2013; Bergelson and Aslin, 2017). We jointly addressed the priorities of sampling

continuous time as well as identifying and excluding extended silences from annotation. First, we inserted silence into any period of a .wav file when the LENA™ digital language processor had been turned off during the day, in order to preserve continuous time. Next, we protected families' privacy by replacing with silence any portions of a .wav file that caregivers noted as private or outside of their home. We then automatically identified sections of the .wav file that fell below a decibel threshold (−22 dB relative to the peak amplitude for that recording) for at least 3 min. This criterion was informed by previous research (Bergelson and Aslin, 2017) and verified through testing on pilot data. Finally, we manually identified any brief sounds (under 3 min) that interrupted two otherwise adjacent periods of silence (e.g., the baby sneezed while napping) as well as any extended periods (at least 10 min) of highly regular sound (e.g., a white noise machine on during the baby's nap). These pre-processing steps generated one .txt file per recording that was read into ELAN to show the start and end times of to-be-annotated sections of the recording. Pre-processing yielded roughly 270 h of to-be-annotated data, which was 0.42 of the total recorded data. This reduction was expected due to the typical duration of sleep and mix of other activities for infants in this age range (Galland et al., 2012). Overall, we integrated the realities of unknown or sparse base rates of everyday music with theory-irrelevant portions of infants' days to settle our sampling scheme of continuously annotating pre-processed everyday audio recordings.

PRINCIPLE 3: FINEST-GRAINED DEFENSIBLE TIMESCALE OF OBSERVED PHENOMENA IN MANY HOURS OF EVERYDAY LIFE

With continuously recorded daylong data, it is theoretically possible to quantify rhythms at every timescale from yoctosecond (i.e., one septillionth of a second) to full-day (i.e., one 24-h time period). Practically, temporal resolution is constrained in part by the device that recorded the everyday data. For example, the LENA™ digital language processor decompresses recorded sound at a resolution of 16 kHz (Ford et al., 2008), which means that milliseconds (not yoctoseconds) is the finest-grained available timescale. Beyond device sampling rates, it is widely acknowledged that “the hardest problem in dealing with time is determining the appropriate sampling intervals” (Adolph, 2019, p. 191). For example, should one manually annotate the presence or absence of music per every millisecond, second, minute, or hour throughout continuously recorded days? Most considerations point toward a principle of sampling finer- rather than coarser-grained.

Evidence about the duration of individual episodes of one's annotation target should inform decisions about which timescale(s) to manually annotate. For example, we know that consequential behaviors in many domains are brief and last on the order of milliseconds to seconds (e.g., Adolph et al., 2012; Warlaumont et al., 2014; Suarez-Rivera et al., 2019; Mendoza and Fausey, 2021a). Because instances of these brief

behaviors are not uniformly available over the full duration of an extended recording, sampling much more coarsely than their individual durations may distort prevalence estimates. For example, suppose that individual music bouts persist for seconds (not hours) and manual annotation designed to detect music within a daylong recording identifies the presence or absence of music per hour. Suppose that at least one music bout occurs within every hour, yet very few bouts persist for its entire hour. Hourly annotation sampling would yield the (distorted) conclusion that music is constant throughout the day. Note that because the durations of many everyday behaviors are currently unknown, discoveries about many temporal rhythms would advance developmental theory even if it is possible to code even finer-grained. For example, everyday rhythms annotated “per minute” are much finer-grained than “per day,” “according to retrospective caregiver report,” or “per year, theorists assume.”

Relatedly, many devices sample less frequently than once per second (e.g., Mehl, 2017; Casillas et al., 2020) in order to achieve extended battery lives. Annotators may also sample more coarsely than devices, respecting properties of human perceivers’ temporal resolution or rates of environmental change. For example, researchers annotated everyday egocentric visual rhythms at 1/5 Hz from recordings captured at 30 Hz in order to make initial discoveries about the prevalence of faces and hands (Fausey et al., 2016). “Down-sampling” is sometimes used to describe such schemes, yet the resulting annotations offer theorists finer-grained insights about everyday ecologies than extant evidence. The priority is not to describe every phenomenon at the finest-grained timescale of *any* observed phenomenon, but rather at a timescale that advances theory by annotating at a temporal resolution hypothesized to be theory-relevant yet currently unknown for the target behavior.

Manually annotating many hours of everyday life also creates opportunities to discover relationships among multiple timescales. The reality of multiple nested timescales minimizes pressure to pick “the” “right” timescale because any single timescale is limited in its explanatory value for developmental change when considered in isolation (Thelen and Smith, 1994; Spencer et al., 2011). Insights about multi-scale structure could arise by aggregating across distinct investigations, or it could be the goal of a single annotation effort. For example, Allan Factor captures hierarchical clustering and can be quantified by annotating at a finer grain and then aggregating across increasingly coarser grains (Abney et al., 2017b; Falk and Kello, 2017). Recent tutorials provide theorists with additional inspiration and considerations about structure at multiple timescales of everyday experiences (Xu et al., 2020).

Generally, it is possible to aggregate from finer- to coarser-timescales without additional annotation, but not the reverse (Adolph, 2019). Finer-grained annotations also make for everyday datasets that are maximally useful as training and evaluation sets for developing automated algorithms to detect everyday behaviors (e.g., Räsänen et al., 2019). Theorists therefore maximize potential for insights for themselves, and for others upon sharing their annotations, by annotating the finest-grained defensible timescale (see also Principle 2). One constraint that places a bound on the finest-grained

defensible timescale is inter-rater reliability. For example, even if a phenomenon varies from 1 millisecond to the next, two annotators may reach similar descriptions only at the timescale of seconds. Designing increasingly laborious annotator training procedures in attempts to achieve reliable finer-grained annotations often yields diminishing returns and so is not feasible (see Principle 4). Other feasibility constraints like personnel time can be managed by strategically structuring multiple passes of annotating everyday data. For example, musical features, voices, and tunes are nested within music bouts (Principle 1). By first annotating temporal onsets and offsets of music bouts at a finer-grained timescale, subsequent passes of judging the presence/absence of features (e.g., “vocal”) and identities (e.g., “Itsy Bitsy Spider”) per bout can yield temporal information without annotators having to also spend person hours marking onsets and offsets of the features and identities. Thus, theorists can optimize a suite of theoretical and practical considerations to annotate the finest-grained defensible timescale of their target phenomena in many hours of everyday life.

Implementation

Figure 1 shows the OSF components with multiple resources relevant for implementing this principle. One illustrative file is “1_MusicBouts_SecMidnight.R” (<https://osf.io/cr2mt>), which smooths everyday music annotations from native ELAN milliseconds into seconds.

Annotating Music Bouts at the Milliseconds Timescale

Little was known about the duration of individual instances of music in infants’ everyday ecologies, so we lacked robust empirical evidence to motivate a timescale for detecting everyday music bouts. We initially annotated pilot recordings at the 5-min timescale, following related manual annotation schemes for efficiently sweeping through many hours of everyday life (e.g., Weisleder and Fernald, 2013). These efforts readily revealed that everyday music bouts were much briefer than 5 min. Thus, we capitalized on ELAN’s native timescale of milliseconds for continuously annotating audio recordings.

Smoothing Annotated Music Bouts to the Seconds Timescale

We smoothed ELAN annotations to the seconds timescale for two reasons. First, some evidence suggested that infants would encounter playsongs and lullabies (Trehub and Schellenberg, 1995; Trehub and Trainor, 1998) whose composed renditions last for seconds not milliseconds (e.g., a typical rendition of “Itsy Bitsy Spider” takes ~17 s). Second, though ELAN’s default timescale is milliseconds, we did not train annotators to obsess about millisecond precision in music bout onsets and offsets which would have required listening and re-listening with unclear payoff for initial discoveries. Thus, we smoothed the atheoretical native resolutions of LENATM and ELAN to a timescale of our observed phenomenon.

To format music bouts data into the seconds timescale, we exported the annotated data from ELAN with one row per music bout indicating its onset and offset times in milliseconds and seconds. We inclusively rounded the ELAN onset and offset times to the nearest second. We expanded the ELAN data into a timeseries of seconds starting at 0 (midnight) and continuing for 129,600 s (i.e., a 36-h time span), to achieve a shared dataframe across recordings that accommodated a small number of recordings that were recorded later in the day. We populated each second (row) in which an annotator identified music with a “1” and the remaining with “0.” If two consecutive music bouts were separated by <1 s as annotated in ELAN, then they were merged into one music bout in this timeseries. In this way, each annotator’s data were transformed into a common format: a .csv file with 129,600 rows representing each second in a 36-h period starting at midnight of the recorded day. We analyzed everyday music at this timescale of seconds.

Merging Annotations of Musical Features, Voices, and Tunes

In subsequent annotation passes, annotators identified the features, voices, and tunes in each music bout ($N = 4,798$ bouts). These additional annotations were per bout, obviating any need for further timescale operationalizations.

Features, voices, and tunes were originally annotated per music bout. Annotators listened to each previously identified music bout using ELAN and recorded their new annotations in Excel (i.e., one row per music bout with columns for features, voices, and tunes; see OSF for additional details, <https://osf.io/qjpx/>). These annotations were then cleaned (e.g., removed punctuation, checked for typos) and any internal inconsistencies were remedied (e.g., a music bout annotated as “vocal” but without a voice identity; see Mendoza and Fausey, 2021a). Voice and tune identities were then replaced with de-identified labels (e.g., VoiceID1, VoiceID2) in order to protect the confidentiality of participating families.

Annotations were merged into the seconds timeseries. All seconds within a bout inherited any feature, voice, or tune identified within that bout. The disadvantage of this scheme was potential imprecision for bout-internal durations for bouts that had multiple musical features (e.g., “live” and “recorded”), voices (e.g., “Beyoncé” and “Daniel Tiger”), and/or tunes (e.g., “Old MacDonald Had a Farm” and “I’m a Little Teapot”). The advantage of this scheme was savings in person hours (**Supplementary Table 5**, OSF, <https://osf.io/htx57/>).

We discovered that many musical bouts were characterized by a single feature, voice, and tune thus yielding straightforward duration estimates. For discoveries based on estimates derived in part from bouts with multiple features, voices, and/or tunes, we conducted more conservative and more liberal analyses. We discovered similar distributional structure whether we analyzed bouts with only a single feature, voice, and/or tune or analyzed all data that potentially overestimated some feature, voice, and/or tune durations (Mendoza and Fausey, 2021a).

PRINCIPLE 4: TRANSPARENT RATIONALE FOR ASSESSING RELIABILITY OF ANNOTATIONS IN MANY HOURS OF EVERYDAY LIFE

Multiple annotations of the same everyday data should point to the same conclusion about its structure. Considerations for assessing reliability like the kind of variable, study design, and assignment of annotators to samples are relevant when annotating many hours of everyday life. Scaling from practices established using smaller and differently structured datasets, however, sometimes presents challenges with non-obvious solutions. Here, we share a mindset for grappling with these issues and point readers to other resources for specific metrics and calculations (e.g., House et al., 1981; Bakeman and Gottman, 1997; Hallgren, 2012).

Attempting to establish reliability when measuring new constructs, at new timescales, and in immense quantities may raise the blood pressure of theorists accustomed to traditionally short and sanitized behaviors captured in laboratory contexts. In the relatively more wild everyday context, it can be challenging to determine what kind and degree of reliability is “good enough.” As with other efforts at the edge of innovation, theorists should not let the perfect be the enemy of the good. Theorists can integrate extant knowledge with newly encountered realities in order to make a case for productive solutions. Innovation does not license a measurement free-for-all, but rather raises the value of showing due diligence, situating one’s contribution, and transparently sharing each step of the process. Transparency is especially valuable so that other theorists can re-use, aggregate, and over time update practices as new consensus emerges.

The metrics used to assess inter-rater reliability, as well as the proportions of recorded data that are annotated in order to assess reliability, vary widely across empirical endeavors. Theorists may struggle to align reliability habits from literatures with which they are familiar to realities of their everyday data. For example, extended timescales often yield low base rates of target behaviors (e.g., many more seconds without than with music in a 16-h everyday audio recording) as well as distributional details that rarely arise at shorter timescales (e.g., it is hard to smooch a day’s 51 distinct tunes into a traditional 5-min laboratory visit). Theorists should therefore engage in due diligence in order to understand the space of available approaches to assess reliability, particularly with respect to related kinds and timescales of everyday data, and share a summary of their review. We illustrate one example of this process in our **Supplementary Table 3** (OSF, <https://osf.io/htx57/>), which is a review of 32 papers and approaches for assessing reliability of manual annotations of some form of children’s unstructured activity. Such a review is not designed to be exhaustive, but rather helpful in combating failures of imagination about potential metrics, practices, and acceptable thresholds for inter-rater reliability in many hours of everyday data.

We flag two properties of everyday data that often rise in salience as theorists consider reliability and can prompt clarity and revision to other aspects of an overall manual annotation

workflow. First, because nobody re-lives the same second, minute, or hour all day long, data from many hours of everyday life include periods of activity and periods of inactivity. For many infants, naps may be one source of relatively silent periods within a day. Other rising and falling rhythms of target behaviors, due in part to the changing activities of the day (Bruner, 1975; Roy et al., 2015; Montag et al., 2018), can yield low base rates of target behaviors at a daily timescale. Should theorists include or exclude periods of inactivity in their reliability assessments, and does it depend on the source and/or temporal extent of inactivity? This issue is a construct and sampling issue rather than a reliability issue *per se*. Theorists must articulate the extent to which they aim to discover structure that includes naps; if they aim to quantify structure only in infants' waking hours, then periods of naps should not be annotated at all. Similarly, extended periods outside the home can yield acoustic properties that are distinct from most other periods of a day and could therefore be outside the scope of one's central discoveries. Second, reliably annotating everyday data becomes increasingly challenging at ever finer-grained timescales. Pilot annotation efforts that reveal unreliable annotation at one timescale often make a coarser timescale the most defensible (see also Principle 3). Is it still worth it to identify structure at coarser timescales, particularly if this diverges from typical quantifications of related behaviors sampled in more constrained contexts? As noted above, the answer is often yes. In many domains, discoveries of even hourly everyday rhythms would advance knowledge beyond current understanding and also guide future waves of inquiry. Altogether, it is productive to center contributions to developmental theory rather than prioritizing practices (often established in contexts of "high base rates reliably coded at the millisecond timescale") that may not scale to the everyday context.

Another source of potential indecision on the way to reliable manual annotation of many hours of everyday life is establishing the quantitative threshold for "good enough." In the absence of formal consensus, transparency is the way forward. Three strategies to arrive at an achievable and productive reliability threshold include (1) identify typical ranges of reliability via systematic review of related everyday phenomena, (2) identify the current state of algorithm-human concordance, and (3) identify the set (if not all) of captured data that can be reliably annotated.

Systematic review of related evidence (as in **Supplementary Table 3**, OSF, <https://osf.io/htx57/>) calibrates typical ranges of reliability. The achievable reliability ceiling in everyday data may be lower than in laboratory data due to lower signal-to-noise ratios arising for various reasons. Systematic reviews are themselves publishable as incremental contributions to growing literatures. Another strategy to help calibrate one's reliability threshold is to identify concordance between commonly accepted algorithmic estimates and human annotation (e.g., Cristia et al., 2021). If one's human-human annotation concordance exceeds algorithm-human concordance, then one's annotation scheme ranks favorably compared to insights based on algorithmically detected patterns. Finally, theorists can plan to analyze only those portions of their data that are reliably annotated. Multiple annotators can judge individual episodes of everyday behavior, and then only those

episodes that are annotated identically by multiple annotators are analyzed (e.g., Fausey et al., 2016; Cychosz et al., 2021). With this approach, one need not drop an entire project because some of the data are difficult to annotate and contribute to a low "overall" reliability. The resulting reliably annotated dataset is often orders of magnitude larger and more diverse than other data available to advance developmental theory. Note that if the bulk of the data are difficult to reliably annotate, then theorists should revisit Principle 1 in order to design an annotation scheme that is better matched to everyday instantiations of their construct. Overall, theorists can transparently situate their contribution as "good enough" with respect to extant knowledge about their target phenomenon.

Two further dimensions of assessing reliability when annotating many hours of everyday life lead theorists to confront tension between scientific rigor and daunting personnel effort. First, should annotators identify everyday behaviors by continuously listening to or watching recordings, or could they instead annotate pre-segmented clips? Second, should multiple people annotate all data, or could some smaller proportion of the data be submitted to reliability computations? Continuous annotation is necessary when one's primary research question is to discover the durations of everyday behaviors. Continuous annotation can also make for higher reliability when one's goal is to detect repetitions of like kind (e.g., the same tune sung in the morning and in the afternoon) by maintaining available context cues from a particular family. Tools like ELAN and Datavyu make continuous coding reproducible. Under certain sampling or signal-to-noise scenarios (see Principle 2 and above), pre-segmented clips are justified and efficient. Full, and not partial, reliability is most productive when implementing an annotation scheme for the first time or in a very new context (e.g., everyday music, Mendoza and Fausey, 2021a). When annotating behaviors with wide consensus about their operationalization (e.g., utterances, words), partial reliability suffices. For partial reliability protocols, best practice is to annotate partial datafiles rather than partial datasets (e.g., 20% of each infant's recording instead of 20% of recordings; Adolph et al., 2019).

From rationale to implementation, transparency has never been easier. Increasingly, systematic reviews and meta-analyses are available (e.g., Ganek and Eriks-Brophy, 2018). Sharing one's own due diligence is straightforward (e.g., Open Science Framework). Visualizations of data together with figure captions that highlight relevant reliability can also be helpful in bridging expertises across scholarly communities (e.g., Figure 3 in Mendoza and Fausey, 2021a). Taking advantage of shared protocols can also save theorists from reinventing every aspect of a workflow (e.g., Adolph et al., 2019; Soderstrom et al., 2021). Over time, as more theorists tackle annotating many hours of everyday life in order to advance theories of developmental change, new consensus will emerge. Each theorist contributes to this consensus by making their rationale for assessing reliability transparent.

Implementation

Figure 1 shows the OSF components with multiple resources relevant for implementing this principle. One illustrative

file is “3_IRR_Tunes_Part2.R” (<https://osf.io/jgw57/>), which is used to calculate Tschuprow’s T for assessing contingency between multiple annotators’ distributional structure of everyday musical tunes.

As mentioned, we reviewed relevant literature in order to calibrate our approach to reliability in the new endeavor of quantifying theory-relevant properties of everyday music in infancy and we shared this review (**Supplementary Table 3**, OSF, <https://osf.io/htx57/>). Annotators continuously annotated daylong recordings, skipping silent portions (Principle 2), and so reliability computations considered these annotations. Because this was the first time anyone had quantified music and its features, voices, and tunes in many hours of everyday life, each annotation pass of each recording was fully annotated by two independent annotators. We calculated a Pearson correlation coefficient to assess reliability of annotated music bouts. For each annotated musical feature, we calculated proportion agreement at the level of music bouts. For the annotated voice and tune identities, we calculated Tschuprow’s T , because this metric allowed us to compare two sets of nominal data with potentially different numbers of unique categories in each set of manual annotations (e.g., if Annotator 1 listed 26 unique voices and Annotator 2 listed 23 unique voices). We determined Tschuprow’s T at the level of music bouts for annotated voice identities and tune identities. For all of these metrics, we used a reliability threshold of 0.90 because this was squarely within the range of previously reported values. Inter-rater reliability was high for all annotations (Mendoza and Fausey, 2021a). To facilitate cumulative science, we shared our data, our scripts for computing reliability, and detailed instructions about our reliability procedure (OSF).

PRINCIPLE 5: ACTIVE AND REPRESENTATIVE TRAINING PROTOCOL FOR ANNOTATING MANY HOURS OF EVERYDAY LIFE

Every annotator must learn the detailed procedure for manual annotation and execute it reliably. The challenge, then, is how to train initially naïve annotators. Traditionally, scholars have lacked robust guiding information about how to design a training protocol for reliably annotating a complex phenomenon in many hours of everyday life. Encouragingly, this is rapidly changing and we contribute some further resources here (Casillas et al., 2017; Adolph et al., 2019; Soderstrom et al., 2021; see also Ramírez-Esparza et al., 2014; Belardi et al., 2017).

The task of manually annotating the full duration of a daylong recording requires annotators to maintain a very high level of attention to detail across many, many hours of work. Any training protocol must successfully prepare and evaluate annotators for this challenge. The principle, then, is to create an active and representative training protocol. A first phase that emphasizes active learning serves to train general skills, with annotators actively practicing generating annotations, making predictions, and asking questions. Annotators are then evaluated on their ability to annotate recordings that match the real data

in both total duration and content in a second phase designed to reveal annotators’ potential lapses in attention and memory across many hours.

Implementation

Figure 1 shows the OSF components with multiple resources relevant for implementing this principle. One illustrative file is “1_MusicBouts_Coding Training_Script.pdf” (<https://osf.io/xd85u/>), which shows how to conduct a training session for annotators learning to annotate music bouts.

Offer Comprehensive Training Sessions

In our procedure, trainees actively participated in two separate training sessions, led by an expert annotator: a 1-h session on how to annotate music bouts and a 2- to 3-h session on how to annotate musical features, voices, and tunes. For each, we started with a brief overview of the study, to help trainees understand what they were listening to and why. Then, we explained the training goal: if their annotations of a full-length training recording matched at least 0.90 with those of an expert annotator, then they would be considered a “trained annotator” and they could annotate real data. Several key features of these training sessions encouraged active learning: (1) Trainees got hands-on practice navigating the server that hosts the data, setting up annotation files, and creating annotations. This was helpful because trainees’ prior computer use varied widely. (2) Trainees reviewed each slide of the coding manual and listened to every audio example. We asked them to generate their own ideas about why each example was included. This, in combination with feedback from the expert annotator, helped trainees learn what they were supposed to annotate. (3) We created step-by-step instructions with screenshots for how to complete every step of the annotation process. These “how-to” documents are rich sources of information for annotators that reduced their cognitive load for completing this work. (4) We provided explicit instructions about taking breaks, working independently, and not multitasking, boosting annotators’ ability to maintain high-quality work across many hours. (5) We encouraged trainees to ask questions throughout this training session. We aimed to make it clear that this type of work requires high attention to detail and also that we would provide them with a lot of support.

Design Representative Training Recordings

We used six daylong recordings from pilot data as training recordings. An expert annotator identified music bouts in all six recordings and then annotated features, voices, and tunes in the music bouts of three of these recordings (**Supplementary Table 4**, OSF, <https://osf.io/htx57/>). We designed training recordings that contained a range of targets that annotators would need to identify. The three recordings for annotating music bouts each had multiple instances of music, a mix of sounds that were easier and harder to identify as music, and musical sounds that occurred in one or multiple music bouts. The three recordings for annotating features, voices, and tunes each contained a wide range of musical voices and tunes in all combinations of features and included music bouts with multiple voices and/or multiple tunes.

Require Trainees to Annotate and Meet Criterion on a Representative Training Recording Prior to Annotating Real Data

The basic skills needed to identify music bouts and to annotate features, voices, and tunes in a 5-min recording are the same as those necessary for annotating a daylong audio recording; the challenge is endurance. We required trainees to practice annotating a full-length training recording in order to assess the extent to which they could maintain high-quality annotating across the entire duration of a daylong audio recording. Trainees annotated separate recordings for music bouts and for musical features, voices, and tunes. We used the same criteria and procedures for assessing reliability as described in Principle 4, with trainees' annotations compared to the expert's annotations. If a trainee failed to reach the 0.90 criterion on their first training recording, then they received feedback, practice, and further training. They could then annotate up to two additional training recordings. If they failed to reach criterion on all three training recordings per annotation pass, then they never annotated real data for this project. Roughly three-quarters of trainees met criterion on their first recording, with a handful achieving criterion after two or three recordings. Occasionally, a trainee failed to reach criterion and/or decided to stop working on this project prior to completing the training.

PRINCIPLE 6: PROCEDURAL PRIORITY ON SELECTIVE ATTENTION WHEN ANNOTATING MANY HOURS OF EVERYDAY LIFE

In daylong audio recordings of everyday environments, there is a lot to notice in the complexity of real life (Xu et al., 2009). Annotators are tasked with identifying a specific phenomenon, such as music, among a mix of many everyday sounds, including people talking, siblings laughing, dogs barking, dishwashers running, and more. This task presents several challenges. Annotators may encounter multiple, varying forms of a complex phenomenon of interest. Annotators may need to use lots of information in order to identify the phenomenon, such as detailed definitions and multiple audio examples of which sounds should and should not be annotated as music. It may not be possible for annotators to learn and remember all forms of the phenomenon in advance. For example, no single annotator could be expected to recognize every tune and every artist from every genre of music on the radio. With so much information to keep in mind, annotators' attention may drift both in the moment and also over time across a long-term project. The solution to these challenges is to build in practices that support annotators' attention and memory. Thus, designing a procedure that prioritizes selective attention is the principle. Researchers can boost annotators' attention and memory by including (1) distinct annotation passes for annotating one well-defined annotation target at a time, (2) regular review of annotation targets as well as options for searching for and creating annotation labels, and (3) routine quality assurance

checks. These aspects of an annotation procedure reduce the challenges of annotating complex phenomena in many hours of everyday life.

Implementation

Figure 1 shows the OSF components with multiple resources relevant for implementing this principle. One illustrative file is "5_MusicBouts_CheckUpClips_InstructionsToCoders.pdf" (<https://osf.io/cn3ke/>), which shows one example of step-by-step instructions to annotators as well as check-up clips used for routine quality assurance checks.

Include Distinct Annotation Passes for Distinct Annotation Targets

To reduce the challenge of identifying the many forms of music, we implemented five separate annotation passes, each with one distinct target: music bouts, live and/or recorded features, vocal and/or instrumental features, voice identities, and tune identities. This procedure prioritized selective attention by requiring annotators to focus on one well-defined annotation target (i.e., one property of music) at a time, thereby minimizing the cognitive burden for annotators.

Require Regular Review of Annotation Targets

We provided annotators with manuals that contained a lot of information and audio examples for identifying the annotation targets of each pass (Principle 1). To boost annotators' memory for this information, annotators reviewed the relevant section of the manual for their current pass at the beginning of each work session. This helped annotators to keep the definitions and examples of music, features, voices, or tunes fresh in their minds. It also helped them to transition from whatever activity they were doing before their work session into the annotation task, thus enhancing selective attention and minimizing divided attention (e.g., sending e-mails and/or working on coursework). Annotators could also return to the manual at any point during their work sessions, which further reduced the amount of information they needed to hold in their working memory while annotating.

Allow Searching for and/or Inventing Labels

There are clear limits to annotators' knowledge of and memory for all potential forms of music and musical features, voices, and tunes. For example, an annotator might recognize a radio voice as Beyoncé but not know the name of the tune. Our procedure included two elements that enabled annotators to increase their own knowledge of the many forms of music: (1) Annotators completed a one-time media review prior to annotating the data for musical features, voices, and tunes. This review familiarized annotators with the wide range of musical sounds likely to occur in infants' everyday environments (i.e., TV shows, music, and toys created for children and for adults). It also reminded them that they would likely hear musical sounds from sources they have not personally encountered before (e.g., a children's TV show that they had not seen) and that they should still strive to identify the specific voices and tunes therein. Note that

the examples in this media review were from Western culture, intentionally selected from the cultural context in which the participating families lived. (2) Annotators used the internet to search for voice and tune identities when they heard a musical sound that they could not immediately recognize. They were not allowed to use any song-identifying software (e.g., Shazam) that would directly access the raw audio recordings (i.e., confidential data). They were also not allowed to consult any human resource since this could violate the independence of their annotation and/or compromise data confidentiality. In addition to searching for existing labels, annotators were allowed to invent their own open-ended labels for voice and tune identities if they could not determine the specific identity for standard tunes (e.g., “upbeat pop song”) or if someone in the recording invented a tune on the spot (e.g., “parent’s toes song”). License to invent labels helped annotators avoid perseverating on never-ending searching or second-guessing and released attention to tackle subsequent annotations.

Build in Routine Quality Assurance Checks

Across a project, annotators might pay less attention to the detailed definitions for music and musical features, voices, and tunes. For example, an annotator could at some point start to judge a parent’s vocal car sound effects as music, even though these kinds of sound effects are explicitly listed as not music in the manual. To avoid this kind of attentional drift, annotators completed routine quality assurance checks. These checks consisted of manually annotating one “check-up clip” after every two daylong audio recordings annotated. For music bouts, each check-up clip was either one 20-min segment or two 10-min segments selected from pilot audio recordings. The expert annotator manually annotated each check-up clip. We compared the trained annotator’s manual annotations with those of the expert annotator, using the same procedures for assessing agreement as for the full training recordings (Principle 4), with one exception. Because the duration of check-up clips was short, any single agreement or disagreement (that could be random) carried more weight. So, we adjusted the check-up agreement criterion from $r = 0.90$ to $r = 0.80$. If annotators met this agreement criterion, then they resumed annotating real-data audio recordings. If not, then they were given up to two more check-up clips to annotate. If their annotations of the second or third check-up clip met the agreement criteria, then they returned to annotating real data. If they did not reach the agreement criteria on any of the three check-up clips, then they did not annotate any further real data and their annotations for their two most recently annotated recordings were replaced. We implemented the same check-up clip procedure for manually annotating features, voices, and tunes. These check-up clips each consisted of 10 music bouts selected from pilot recordings annotated by the expert annotator. We used the same agreement criteria as for the full training recordings and the same logic for determining if an annotator should continue to annotate real-data recordings. Overall, no annotator’s manual annotation drifted to the point that they were removed from the project.

PRINCIPLE 7: INCLUSIVE PHYSICAL, SOCIAL, AND PROFESSIONAL COMMUNITIES FOR ANNOTATORS OF MANY HOURS OF EVERYDAY LIFE

Manually annotating many hours of everyday life is not an easy task. It takes a lot of time. Annotators spend long hours working at their computer stations. The bulk of the work must be done independently, so annotators may feel isolated or like they are a cog in a machine. Annotators must sustain high levels of focus in order to detect specific targets that may occur infrequently. This makes the task simultaneously cognitively demanding and boring. The key challenge is to maintain motivation among annotators so that they continue to generate high-quality annotations throughout a long-term project. Creating a healthy community is the principle. Recruiting a large, diverse group of annotators creates an inclusive community. Encouraging annotators to work as a team helps them develop a sense of belonging and feel invested in the work. Providing opportunities for annotators to build skills and to receive mentorship from senior colleagues adds to the value for annotators, keeping them engaged in the work. Adding in fun activities recognizes annotators’ humanity, increasing their enthusiasm to be actively involved. Setting up a physical workspace with varied ways to work comfortably makes annotators ergonomically happy. Annotators who work as part of inclusive physical, social, and professional communities are more likely to stay and to do high-quality work. Avoiding high team turnover is especially important when training procedures require investing roughly 25 h per person.

Implementation

Figure 1 shows the OSF components with multiple resources relevant for implementing this principle. One illustrative file is “2_LearningLab_Bingo_Winter2018.png” (<https://osf.io/x4gd8/>), which is one example of a lab practice designed to promote community.

Recruit a Large and Diverse Team

In the University of Oregon Learning Lab, we do not require undergraduate students to be psychology majors, to have research experience, or to have completed specific coursework prior to applying for a research assistant (RA) position. These practices reduce some systemic barriers for institutionally underrepresented students in academia to become directly involved in research, actively promoting equity and inclusion. For this project, we also did not require students to have prior formal or informal musical training. Our team of music annotators (**Figure 2**) had varied majors, including psychology, music, computer science, physics, sociology, linguistics, and human physiology. Their music experience ranged from none to lifelong. We found this mix to be beneficial because they collectively provided a wide range of insights and observations. Manual annotation research combined with our approach to building a team is well-suited to diversifying the scientific workforce.



FIGURE 2 | Photos of UO Learning Lab team members who manually annotated music and musical features, voices, and tunes, including (top row, left to right): Dr. Caitlin Fausey (PI), Dr. Jennifer Mendoza (doctoral student at the time), Catherine Diercks (lab manager), Christine White (lab manager), and 35 research assistants (left to right, Row 2: Hitomi Tanizawa, Josh Mabry, Vinitha Gadiraju, Emma Salmon, Adeline Fecker, Helen Rawlins, Madison Edgar, Sabrina Haskinson, Kayla Figone, Row 3: Aiko Luckasavage, Samuel Hickman, Melissa Lattig, Erin Batali, Katie Mickel, Sophie Cohen, Thorin Faulk, Jennifer Lowery, Row 4: Jayne Coles, Cayla Lussier, Amanda Powell, Kyra Wilson, Jordyn Mons, Grace Floyd, Juliette Tisseur, Arie Markowitz, Row 5: Brittany Brann, Mitchell Passadore, Allysia Rainey, Natalie Draga, Liam Green, Melissa Berg, Kelly Woltjer, Rachel Ward, Jewel Montes, Keelan Paroissien-Arce).

Meet Often in Varied Configurations

Our team actively participates in multiple weekly lab meetings, each designed to advance our scientific research and to promote professional development. Meeting with varied configurations of lab personnel (in person or virtually) provides opportunities for lab members to build different skills. For this project, annotators attended project team meetings, led by Mendoza, where they discussed the ongoing music annotation work, asked questions, and shared observations about the data (limiting specifics to preserve annotator independence). They built skills for project management by collectively reviewing progress, setting concrete goals, and prioritizing weekly tasks. During our full-lab meeting, all lab personnel participated in a mix of science-skill-building activities, including discussing empirical studies, giving elevator pitches about our work, and using statistical computing tools (e.g., R & Python). RAs steered these meetings, voicing their ideas and questions. Lastly, during small senior personnel meetings, Fausey met with graduate students, lab managers, and select senior RAs. Trainees took the lead, asking questions, collectively problem solving, and soliciting feedback about their research, thus supporting senior personnel to develop more advanced science skills. By holding each of these different meetings on a weekly basis, we lowered barriers to identifying problems and accelerated finding solutions.

Include Activities That Recognize the Humanity of the Team

To build a sense of community, we regularly include tasks designed to humanize the experience of working in a research lab. At the beginning of each meeting, everyone shares a fun fact about themselves. When large teams regularly use the lab space, Mendoza and Fausey frequently work in the lab, intentionally creating opportunities to talk with lab personnel. Each term we play lab bingo, with bingo cards filled with science tasks (e.g., made a plot, asked a question, used R, called a family) and we have bingo prizes. We make up science raps and songs, both for entertainment and to boost our learning. Having fun and being part of a community is motivating; lab personnel gain the sense that they matter, that their work matters, and that their work affects other people in the lab.

Design a Physical Workspace to Promote Well-Being

We created a workspace that allowed for many RAs to work simultaneously so they would not feel isolated. We had computers dedicated for annotators to use while annotating music data. We also maintained a large assortment of chairs (including yoga balls) and an adjustable standing desk for annotators to use. This mix of furniture helped keep lab personnel ergonomically happy and minimized the potential for

repetitive stress injuries from working long hours at a computer (Tomba et al., 2010). Lonely annotators who are in physical pain will generate low-quality work and ultimately leave. Thus, creating a physically and socially inclusive work environment was critical to the success of our annotation team.

Using these multiple strategies, we created an inclusive physical, social, and professional community and provided lab personnel with rich educational experiences. Our healthy, positive working environment supported annotators to conduct an estimated total of 6,400 h of manual annotation (**Supplementary Table 5**, OSF, <https://osf.io/htx57/>). Annotators understood that we valued their contributions to the team. They also recognized that we were supporting their professional development. In addition to direct experience conducting research, they gained knowledge and skills that would help prepare them for a wide range of future positions both within and outside developmental science.

DISCUSSION

Insights into infants' everyday ecologies are increasingly available to developmental theorists thanks to the combination of wearable technologies to record these ecologies (de Barbaro, 2019), infrastructure to share the everyday recordings (MacWhinney, 2000; VanDam et al., 2016; Gilmore et al., 2018), and protocols to facilitate detecting structure in these everyday samples (Adolph et al., 2019; Soderstrom et al., 2021). Rigorously quantifying everyday ecologies advances theories of developmental change by centering tasks, timescales, and trajectories that are not discoverable in traditional laboratory protocols (Dahl, 2017; Rogoff et al., 2018; Smith et al., 2018; Frankenhuus et al., 2019; de Barbaro and Fausey, in press). One exciting and daunting frontier is to quantify everyday opportunities for learners to attend, encode, retrieve, and integrate experiences over not just one but many timescales. Here, we articulated a framework with an eye toward optimizing multi-scale dividends upon investing considerable resources in manually annotating many hours of everyday life. We encourage theorists to jump into the endeavor of quantifying everyday ecologies in order to make discoveries about the experiences upon which development may depend.

Importantly, everyday ecologies vary across the world's communities. Cross-cultural variation is evident in infants' opportunities to encounter child-directed speech (Casillas et al., 2020), sing with caregivers (Trehub and Schellenberg, 1995), move and explore (Rachwani et al., 2020), and more. Multiple levels of context organize experiences (Rowe and Weisleder, 2020); quantifying everyday ecologies across variation in these contexts will advance theories about multiple pathways of developmental change. This ambitious goal is attainable in part by annotating existing corpora that span the world's cultures (e.g., Benetti and Costa-Giomi, 2019; Bergelson et al., 2019) as well as mindfully sampling and annotating more everyday ecologies over time. We highlight that Principles 1 and 6 may prove especially helpful across distinct annotation endeavors, with research teams investing effort in iterative pilot annotations in order to arrive at constructs that are meaningful within specific communities as well as training annotators with representative recordings.

Relatedly, accelerating the breadth and pace of discoveries is also more likely with an ever more diverse and inclusive community of scholars. Manually annotating existing corpora is one research activity that is amenable to contributions across researchers who have varying expertises, working environments, cultural contexts, and resources. Such diversity is also deeply necessary in order to minimize biases in operationalizing everyday behaviors as they arise in many contexts (e.g., Cychosz et al., 2020a). Aggregating contributions from many individuals and teams means that smaller efforts cumulate to larger insights, making team science well-matched to the challenge of annotating many hours of infants' everyday lives (see also Cychosz et al., 2021). Frameworks designed to address issues that arise in research that is distributed over time and teams include co-authorship and contributorship models (Holcombe, 2019; Moshontz et al., 2021), pre-registration of secondary data analyses (Van den Akker et al., 2019), and protocols devised for widespread use coupled with practical tutorials to support incremental contributions (Soderstrom et al., 2021).

Resources like HomeBank (VanDam et al., 2016), Databrary (Gilmore et al., 2018), and Open Science Framework (<https://osf.io/>) are vital to maintain and expand because they make it possible for theorists to transparently make progress collectively. These are living repositories, potentiating new discoveries about human development through curation of more and different data over time. Notably, multiple funding agencies helped launch these repositories and dedicate specific grant mechanisms to support secondary data analyses at multiple scales (e.g., NIH R03, NSF SBE HNSD-I). Continued investment and diverse engagement will maximize the value of these collective treasures and propel developmental science forward.

Overall, contributions of many kinds will be necessary to build a diverse and cumulative science of everyday ecologies. For discussion of practical tradeoffs facing any individual theorist—including rapid vs. delayed theoretical gratification, going it alone vs. collaborating, and sampling selectively vs. exhaustively—see de Barbaro and Fausey (in press). One sign of productive manual annotations at scale will be its demise after theorists have used annotated everyday datasets to successfully train algorithms to automatically detect theory-relevant behaviors in the hubbub of everyday sights, sounds, and more. We are currently very far from this goal in most domains and there may be no surer way out than through. Manually annotating many hours of infants' everyday lives is likely to spur innovation not only in theories of developmental change but also in the tools used for future discoveries.

Quantifying everyday ecologies can inspire theorists to pursue hypotheses that might not arise from other sources (see also Nastase et al., 2020). For example, instead of presenting learners with input distributions inspired by traditional laboratory instantiations of consistent amounts of multiple category instances distributed evenly over time, learning theorists might instead appreciate the striking prevalence of non-uniform content and temporal distributions in everyday ecologies (e.g., Smith et al., 2018; Mendoza and Fausey, 2021a,b) and test hypotheses about the consequences of these distributions (e.g., Casenhiser and Goldberg, 2005; Carvalho et al., 2021). Manipulating training regimes for

both human learners and models, with parameters shown to be plausible in everyday ecologies, will bring developmental theory closer to meeting longstanding goals of jointly modeling the input and its impact (e.g., Smith and Slone, 2017). Quantifying distributions of everyday parameters encountered across learners will also inspire new routes to understanding individualized developmental pathways by combating failures of imagination due to traditional one-size-fits-all training protocols (e.g., Thelen and Smith, 1994; Samuelson, 2021). All of these potentially dramatic expansions to future hypothesis testing are within reach of any theorist making use of everyday corpora.

The current moment in developmental science is full of opportunities for game-changing discoveries by taking advantage of methods that scale beyond traditional laboratory experiments. For example, developmental theorists can now implement experiments beyond the reach of only their local community (e.g., ManyBabies, Frank et al., 2017; Lookit, Scott and Schulz, 2017). New tools enable theorists to aggregate across large bodies of evidence (e.g., MetaLab, Bergmann et al., 2018). Quantifying everyday ecologies similarly scales beyond traditional contexts and timescales available to ground theories of development. The framework presented here can support theorists as they embark on efforts to annotate many hours of infants' lives en route to discovering more about the experiences available to drive experience-dependent change.

DATA AVAILABILITY STATEMENT

In accordance with family consent, audio recordings and extracted music clips are available on HomeBank (doi: 10.21415/T5JM4R; doi: 10.21415/T47D-5K51). Study materials, behavioral coding manuals, numerical

data, and analysis code are available on Open Science Framework (doi: 10.17605/osf.io/eb9pw).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board, University of Oregon. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

JM and CF contributed to all aspects of manuscript preparation. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded in part by a grant from the GRAMMY Museum® to CF.

ACKNOWLEDGMENTS

We thank Catherine Diercks and Christine White for contributing to data collection, team management, and open science material preparation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://osf.io/eb9pw/>

REFERENCES

- Abney, D. H., Kello, C. T., and Balasubramaniam, R. (2017a). Introduction and application of the multiscale coefficient of variation analysis. *Behav. Res. Methods* 49, 1571–1581. doi: 10.3758/s13428-016-0803-4
- Abney, D. H., Warlaumont, A. S., Oller, D. K., Wallot, S., and Kello, C. T. (2017b). Multiple coordination patterns in infant and adult vocalizations. *Infancy* 22, 514–539. doi: 10.1111/infa.12165
- Adolph, K., Gilmore, R. O., and Soska, K. (2019). *Play and Learning Across a Year (PLAY) Project - Protocols & Documentation*. Databrary. doi: 10.17910/b7.876
- Adolph, K. E. (2019). An ecological approach to learning in (not and) development. *Hum. Dev.* 63, 180–201. doi: 10.1159/000503823
- Adolph, K. E. (2020). Oh, Behave! *Infancy* 25, 374–392. doi: 10.1111/infa.12336
- Adolph, K. E., Cole, W. G., Komati, M., Garciaguirre, J. S., Badaly, D., Lingeman, J. M., et al. (2012). How do you learn to walk? Thousands of steps and dozens of falls per day. *Psychol. Sci.* 23, 1387–1394. doi: 10.1177/0956797612446346
- Adolph, K. E., Robinson, S. R., Young, J. W., and Gill-Alvarez, F. (2008). What is the shape of developmental change? *Psychol. Rev.* 115, 527–543. doi: 10.1037/0033-295X.115.3.527
- Anderson, H., and Fausey, C.M. (2019). "Modeling non-uniformities in infants' everyday speech environments [Conference Paper]," in *2019 Biennial Meeting of the Society for Research in Child Development*.
- Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdiscip. Rev. Cogn. Sci.* 8:e1373. doi: 10.1002/wcs.1373
- Bakeman, R., and Gottman, J. (1997). *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge: Cambridge University Press.
- Bambach, S., Crandall, D. J., Smith, L. B., and Yu, C. (2016). "Active viewing in toddlers facilitates visual object learning: An egocentric vision approach," in *Proceedings of the 38th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Belardi, K., Watson, L.R., Faldowski, R.A., Hazlett, H., Crais, E., Baranek, G.T., et al. (2017). A retrospective video analysis of canonical babbling and volubility in infants with Fragile X Syndrome at 9–12 months of age. *J. Autism Dev. Disord.* 47, 1193–1206. doi: 10.1007/s10803-017-3033-4
- Benetti, L., and Costa-Giomi, E. (2019). "Music in the lives of American and Tanzanian infants and toddlers: a daylong sampling [Conference Paper]," in *2019 Meeting of the Society for Music Perception and Cognition*.
- Bergelson, E., and Aslin, R. N. (2017). Nature and origins of the lexicon in 6-month-olds. *Proc. Nat. Acad. Sci.* 114, 12916–12921. doi: 10.1073/pnas.1712966114
- Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., and Amatuni, A. (2019). What do North American babies hear? A large-scale cross-corpus analysis. *Dev. Sci.* 22:e12724. doi: 10.1111/desc.12724
- Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., et al. (2018). Promoting replicability in developmental research through meta-analyses: insights from language acquisition research. *Child Dev.* 89, 1996–2009. doi: 10.1111/cdev.13079

- Bolger, N., and Laurenceau, J. P. (2013). *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. New York, NY: Guilford Press.
- Bruner, J. S. (1975). The ontogenesis of speech acts. *J. Child Lang.* 2, 1–19. doi: 10.1017/S0305000900000866
- Canault, M., et al. (2016). Reliability of the language ENvironment analysis system (LENA™) in European French. *Behav. Res. Methods* 48, 1109–1124. doi: 10.3758/s13428-015-0634-8
- Carvalho, P. F., Chen, C. H., and Yu, C. (2021). The distributional properties of exemplars affect category learning and generalization. *Sci. Rep.* 11:1263. doi: 10.1038/s41598-021-90743-0
- Casenhiser, D., and Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Dev. Sci.* 8, 500–508. doi: 10.1111/j.1467-7687.2005.00441.x
- Casillas, M., Bergelson, E., Warlaumont, A. S., Cristia, A., Soderstrom, M., VanDam, M., et al. (2017). A new workflow for semi-automatized annotations: tests with long-form naturalistic recordings of childrens' language environments. *Interspeech* 2017, 2098–2102. doi: 10.21437/Interspeech.2017-1418
- Casillas, M., Brown, P., and Levinson, S. C. (2020). Early language experience in a Tzeltal Mayan village. *Child Dev.* 91, 1819–1835. doi: 10.1111/cdev.13349
- Casillas, M., and Cristia, A. (2019). A step-by-step guide to collecting and analyzing long-format speech environment (LFSE) recordings. *Collabra Psychol.* 5, 1–21. doi: 10.1525/collabra.209
- Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., and Smith, L. B. (2018). Real-world visual statistics and infants' first-learned object names. *Philosoph. Trans. R. Soc. B Biol. Sci.* 372:20160055. doi: 10.1098/rstb.2016.0055
- Collins, L. M., and Graham, J. W. (2002). The effects of the timing and spacing of observations in longitudinal studies of tobacco and other drug use: temporal design considerations. *Drug Alcohol Depend.* 68, S85–S93. doi: 10.1016/S0376-8716(02)00217-X
- Costa-Giomi, E., and Benetti, L. (2017). Through a baby's ears: Musical interactions in a family community. *Int. J. Commun. Music* 10, 289–303. doi: 10.1386/ijcm.10.3.289_1
- Creel, S. C. (2019). The familiar-melody advantage in auditory perceptual development: Parallels between spoken language acquisition and general auditory perception. *Attent. Percept. Psychophys.* 81, 948–957. doi: 10.3758/s13414-018-01663-7
- Cristia, A., Lavechin, M., Scaff, C., Soderstrom, M., Rowland, C., Räsänen, O., et al. (2021). A thorough evaluation of the language environment analysis (LENA) system. *Behav. Res. Methods* 53, 467–486. doi: 10.3758/s13428-020-01393-5
- Custodero, L. A., and Johnson-Green, E. A. (2003). Passing the cultural torch: musical experience and musical parenting of infants. *J. Res. Music Educ.* 51, 102–114. doi: 10.2307/3345844
- Cychosz, M., Cristia, A., Bergelson, E., Casillas, M., Baudet, G., Warlaumont, A.S., et al. (2021). Vocal development in a large-scale crosslinguistic corpus. *Dev. Sci.* doi: 10.1111/desc.13090. [Epub ahead of print].
- Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., et al. (2020a). Longform recordings of everyday life: ethics for best practices. *Behav. Res. Methods* 52, 1951–1969. doi: 10.3758/s13428-020-01365-9
- Cychosz, M., Villanueva, A., and Weisleder, A. (2020b). *Efficient Estimation of Children's Language Exposure in Two Bilingual Communities*. doi: 10.31234/osf.io/dy6v2
- Dahl, A. (2017). Ecological commitments: why developmental science needs naturalistic methods. *Child Dev. Perspect.* 11, 79–84. doi: 10.1111/cdep.12217
- d'Apice, K., Latham, R. M., and von Stumm, S. (2019). A naturalistic home observational approach to children's language, cognition, and behavior. *Dev. Psychol.* 55, 1414–1427. doi: 10.1037/dev0000733
- Datavyu Team (2014). *Datavyu: A Video Coding Tool. Databrary Project*. New York University. Available online at: datavyu.org (accessed July 28, 2021).
- de Barbaro, K. (2019). Automated sensing of daily activity: a new lens into development. *Dev. Psychobiol.* 61, 444–464. doi: 10.1002/dev.21831
- de Barbaro, K., and Fausey, C. M. (in press). "The promise of mobile sensing for centering tasks and trajectories in developmental theory," in *Mobile Sensing in Psychology: Methods and Applications*, eds M. R. Mehl, C. Wrzus, M. Eid, G. Harari, and U. Ebner-Priemer (Guilford Press).
- Falk, S., and Kello, C. T. (2017). Hierarchical organization in the temporal structure of infant-direct speech and song. *Cognition* 163, 80–86. doi: 10.1016/j.cognition.2017.02.017
- Fausey, C. M., Jayaraman, S., and Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition* 152, 101–107. doi: 10.1016/j.cognition.2016.03.005
- Fausey, C. M., and Mendoza, J. K. (2018a). *FauseyTrio HomeBank Corpus*. doi: 10.21415/T5JM4R
- Fausey, C. M., and Mendoza, J. K. (2018b). *FauseyTrio-Public HomeBank Corpus*. doi: 10.21415/T56D7Q
- Ford, M., Baer, C. T., Xu, D., Yapanel, U., and Gray, S. (2008). *The LENATM Language Environment Analysis System: Audio Specifications of the DLP-0121 (Technical Report LTR-03-2)*. Boulder, CO: Lena Foundation.
- Franchak, J. M. (2019). Changing opportunities for learning in everyday life: infant body position over the first year. *Infancy* 24, 187–209. doi: 10.1111/inf.12272
- Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., et al. (2017). A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy* 22, 421–435. doi: 10.1111/inf.12182
- Frankenhuis, W. E., Nettle, D., and Dall, S. R. (2019). A case for environmental statistics of early- life effects. *Philosoph. Trans. R. Soc. B* 374:20180110. doi: 10.1098/rstb.2018.0110
- Galland, B. C., Taylor, B. J., Elder, D. E., and Herbison, P. (2012). Normal sleep patterns in infants and children: a systematic review of observational studies. *Sleep Med. Rev.* 16, 213–222. doi: 10.1016/j.smrv.2011.06.001
- Ganek, H., and Eriks-Brophy, A. (2018). Language ENvironment Analysis (LENA) system investigation of day long recordings in children: a literature review. *J. Commun. Disord.* 72, 77–85. doi: 10.1016/j.jcomdis.2017.12.005
- Gilkerson, J., Zhang, Y., Xu, D., Richards, J.A., Xu, X., Jiang, F., et al. (2015). Evaluating Language Environment Analysis system performance for Chinese: a pilot study in Shanghai. *J. Speech Lang. Hear. Res.* 58, 445–452. doi: 10.1044/2015_JSLHR-L14-0014
- Gilmore, R. O., Kennedy, J. L., and Adolph, K. E. (2018). Practical solutions for sharing data and materials from psychological research. *Adv. Methods Pract. Psychol. Sci.* 1, 121–130. doi: 10.1177/2515245917746500
- Gómez, R. L., and Edgin, J. O. (2015). Sleep as a window into early neural development: shifts in sleep-dependent learning effects across early childhood. *Child Dev. Perspect.* 9, 183–189. doi: 10.1111/cdep.12130
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor. Quant. Methods Psychol.* 8, 23–34. doi: 10.20982/tqmp.08.1.p023
- Hannon, E. E., and Trainor, L. J. (2007). Music acquisition: effects of enculturation and formal training on development. *Trends Cogn. Sci.* 11, 466–472. doi: 10.1016/j.tics.2007.08.008
- Hannon, E. E., and Trehub, S. E. (2005). Tuning in to musical rhythms: infants learn more readily than adults. *Proc. Nat. Acad. Sci.* 102, 12639–12643. doi: 10.1073/pnas.0504254102
- Hensch, T. K. (2005). Critical period plasticity in local cortical circuits. *Nat. Rev. Neurosci.* 6, 877–888. doi: 10.1038/nrn1787
- Hofferth, S. L., and Sandberg, J. F. (2001). How American children spend their time. *J. Marriage Fam.* 63, 295–308. doi: 10.1111/j.1741-3737.2001.00295.x
- Holcombe, A.O. (2019). Contributorship, not authorship: use CRediT to indicate who did what. *MDPI Publ.* 7:48. doi: 10.3390/publications7030048
- House, A.E., House, B.J., and Campbell, M.B. (1981). Measures of interobserver agreement: calculation formulas and distribution effects. *J. Behav. Assess.* 3, 37–57. doi: 10.1007/BF01321350
- Hruschka, D. J., Medin, D. L., Rogoff, B., and Henrich, J. (2018). Pressing questions in the study of psychological and behavioral diversity. *Proc. Nat. Acad. Sci.* 115, 11366–11368. doi: 10.1073/pnas.1814733115
- Jayaraman, S., Fausey, C. M., and Smith, L. B. (2015). The faces in infant-perspective scenes change over the first year of life. *PLoS ONE* 10:e0123780. doi: 10.1371/journal.pone.0123780
- Kadooka, K., Caufield, M., Fausey, C. M., and Franchak, J. (2021). "Visuomotor learning opportunities are nested within infants' everyday activities. [Conference Paper]," in *2021 Biennial Meeting of the Society for Research in Child Development*.
- Lamont, A. M. (2008). Young children's musical worlds: musical engagement in 3.5-year-olds. *J. Early Childhood Res.* 6, 247–261. doi: 10.1177/1476718X08094449
- Larsen, R. J., and Kasimatis, M. (1990). Individual differences in entrainment of mood to the weekly calendar. *J. Pers. Soc. Psychol.* 58, 164–171. doi: 10.1037/0022-3514.58.1.164

- Lee, D. K., Cole, W. G., Golenia, L., and Adolph, K. E. (2018). The cost of simplifying complex developmental phenomena: a new perspective on learning to walk. *Dev. Sci.* 21:e12615. doi: 10.1111/desc.12615
- MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edn. Mahwah, NJ: Lawrence Erlbaum Associates.
- Margulis, E. H. (2014). *On Repeat: How Music Plays the Mind*. New York, NY: Oxford University Press.
- Mehl, M. R. (2017). The electronically activated recorder (EAR) a method for the naturalistic observation of daily social behavior. *Curr. Dir. Psychol. Sci.* 26, 184–190. doi: 10.1177/0963721416680611
- Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., et al. (2019). Universality and diversity in human song. *Science* 366:eaax0868. doi: 10.1126/science.aax0868
- Mendoza, J. K., and Fausey, C. M. (2018). *MendozaMusic HomeBank Corpus*. doi: 10.21415/T47D-5K51
- Mendoza, J. K., and Fausey, C. M. (2019). *Everyday Music in Infancy*. doi: 10.17605/osf.io/eb9pw
- Mendoza, J. K., and Fausey, C. M. (2021a). Everyday music in infancy. *Dev. Sci.* doi: 10.1111/desc.13122
- Mendoza, J. K., and Fausey, C. M. (2021b). Everyday parameters for daily temporal schedules of music in infancy. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/4n3ef
- Micheletti, M., de Barbaro, K., Fellows, M. D., Hixon, J. G., Slatcher, R. B., and Pennebaker, J. W. (2020). Optimal sampling strategies for characterizing behavior and affect from ambulatory audio recordings. *J. Fam. Psychol.* 34, 980–990. doi: 10.1037/fam0000654
- Mikhelson, M., Micheletti, M., Yao, X., and de Barbaro, K. (2021). “Maternal Mood and Contingency to Infant Distress in Everyday Settings [Conference Paper],” in *2021 Biennial Meeting of the Society for Research in Child Development*.
- Montag, J. L., Jones, M. N., and Smith, L. B. (2018). Quantity and diversity: simulating early word learning environments. *Cogn. Sci.* 42, 375–412. doi: 10.1111/cogs.12592
- Moshontz, H., Ebersole, C. R., Weston, S. J., and Klein, R. A. (2021). A guide for many authors: writing manuscripts in large collaborations. *Soc. Pers. Psychol. Compass.* 15:e12590. doi: 10.1111/spc3.12590
- Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: rethinking the primacy of experimental control in cognitive neuroscience. *Neuroimage* 222:117254. doi: 10.1016/j.neuroimage.2020.117254
- Nielsen, M., Haun, D., Kärtner, J., and Legare, C. H. (2017). The persistent sampling bias in developmental psychology: a call to action. *J. Exp. Child Psychol.* 162, 31–38. doi: 10.1016/j.jecp.2017.04.017
- Ossmy, O., Hoch, J. E., MacAlpine, P., Hasan, S., Stone, P., and Adolph, K. E. (2018). Variety wins: soccer-playing robots and infant walking. *Front. Neurobot.* 12:19. doi: 10.3389/fnbot.2018.00019
- PSID-CDS User Guide (2002). *The Panel Study of Income Dynamics Child Development Supplement User Guide* (2002). Available online at: https://psidonline.isr.umich.edu/CDS/cdsii_userGd.pdf (accessed July 28, 2021).
- Rachwani, J., Hoch, J. E., and Adolph, K. E. (2020). “Action in development: variability, flexibility, and plasticity,” in *Handbook of Infant Development*, eds C. S. Tamis-LeMonda and J. J. Lockman (Cambridge: Cambridge University Press).
- Ramírez-Esparza, N., García-Sierra, A., and Kuhl, P. K. (2014). Look who’s talking: speech style and social context in language input to infants are linked to concurrent and future speech development. *Dev. Sci.* 17, 880–891. doi: 10.1111/desc.12172
- Räsänen, O., Seshadri, S., Karadayi, J., Riebling, E., Bunce, J., Cristia, A., et al. (2019). Automatic word count estimation from daylong child-centered recordings in various language environments using language-independent syllabification of speech. *Speech Commun.* 113, 63–80. doi: 10.1016/j.specom.2019.08.005
- Rogoff, B., Dahl, A., and Callanan, M. (2018). The importance of understanding children’s lived experience. *Dev. Rev.* 50, 5–15. doi: 10.1016/j.dr.2018.05.006
- Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., et al. (2018). Beyond the 30-million-word gap: children’s conversational exposure is associated with language-related brain function. *Psychol. Sci.* 29, 700–710. doi: 10.1177/0956797617742725
- Rovee-Collier, C. (1995). Time windows in cognitive development. *Dev. Psychol.* 31, 147–169. doi: 10.1037/0012-1649.31.2.147
- Rowe, M. L., and Weisleder, A. (2020). Language development in context. *Ann. Rev. Dev. Psychol.* 2, 201–223. doi: 10.1146/annurev-devpsych-042220-121816
- Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., and Roy, D. (2015). Predicting the birth of a spoken word. *Proc. Nat. Acad. Sci.* 112, 12663–12668. doi: 10.1073/pnas.1419773112
- Samuelson, L. K. (2021). Toward a precision science of word learning: understanding individual vocabulary pathways. *Child Dev. Perspect.* 15, 117–124. doi: 10.1111/cdep.12408
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Scott, L. S., Pascalis, O., and Nelson, C. A. (2007). A domain-general theory of the development of perceptual discrimination. *Curr. Dir. Psychol. Sci.* 16, 197–201. doi: 10.1111/j.1467-8721.2007.00503.x
- Smith, L. B., Jayaraman, S., Clerkin, E., and Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends Cogn. Sci.* 22, 325–336. doi: 10.1016/j.tics.2018.02.004
- Smith, L. B., and Slone, L. K. (2017). A developmental approach to machine learning? *Front. Psychol.* 8:2124. doi: 10.3389/fpsyg.2017.02124
- Soderstrom, M., Casillas, M., Bergelson, E., Rosemberg, C., Alam, F., Warlaumont, A., et al. (2021). Developing a cross-cultural annotation systems and metacorpus for studying infants’ real world language experience. *Collabra Psychol.* 7:23445. doi: 10.1525/collabra.23445
- Soderstrom, M., and Wittebolle, K. (2013). When do caregivers talk? The influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments. *PLoS ONE* 8:e80646. doi: 10.1371/journal.pone.0080646
- Spencer, J. P., Perone, S., and Buss, A. T. (2011). Twenty years and going strong: a dynamic systems revolution in motor and cognitive development. *Child Dev. Perspect.* 5, 260–266. doi: 10.1111/j.1750-8606.2011.00194.x
- Suarez-Rivera, C., Smith, L. B., and Yu, C. (2019). Multimodal parent behaviors within joint attention support sustained attention in infants. *Dev. Psychol.* 55, 96–109. doi: 10.1037/dev0000628
- Szymczak, J. T., Jasińska, M., Pawlak, E., and Zwierzykowska, M. (1993). Annual and weekly changes in the sleep-wake rhythm of school children. *Sleep* 16, 433–435. doi: 10.1093/sleep/16.5.433
- Tamis-LeMonda, C. S., Custode, S., Kuchirko, Y., Escobar, K., and Lo, T. (2018). Routine language: Speech directed to infants during home activities. *Child Dev.* 90, 2135–2152. doi: 10.1111/cdev.13089
- Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., and Bornstein, M. H. (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Dev. Sci.* 20:e12456. doi: 10.1111/desc.12456
- Thelen, E., and Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.
- Tomba, E., Dolinschi, R., de Oliveira, C., Amick, B. C., and Irvin, E. (2010). A systematic review of workplace ergonomic interventions with economic analyses. *J. Occup. Rehabil.* 20, 220–234. doi: 10.1007/s10926-009-9210-3
- Trehub, S. E., and Schellenberg, E. G. (1995). Music: its relevance to infants. *Ann. Child Dev.* 11, 1–24.
- Trehub, S. E., and Trainor, L. (1998). “Singing to infants: lullabies and play songs,” in *Advances in Infancy Research*, Vol. 12, eds C. K. Rovee-Collier, L. P. Lipsitt, and H. Hayne (Norwood, NJ: Ablex Publishing Corporation), 43–78.
- Van den Akker, O. R., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P. E., et al. (2019). Preregistration of secondary data analysis: a template and tutorial. *PsyArXiv*. doi: 10.31234/osf.io/hvfmr
- VanDam, M., and De Palma, P. (2019). A modular, extensible approach to massive ecologically valid behavioral data. *Behav. Res. Methods* 51, 1754–1765. doi: 10.3758/s13428-018-1167-8
- VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., et al. (2016). HomeBank: an online repository of daylong child-centered audio recordings. *Semin. Speech Lang.* 37, 128–142. doi: 10.1055/s-0036-1580745
- Vlach, H. A. (2019). Learning to remember words: memory constraints as double-edged sword mechanisms of language development. *Child Dev. Perspect.* 13, 159–165. doi: 10.1111/cdep.12337
- Warlaumont, A. S., Richards, J. A., Gilkerson, J., and Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychol. Sci.* 25, 1314–1324. doi: 10.1177/0956797614531023

- Weisleder, A., and Fernald, A. (2013). Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychol. Sci.* 24, 2143–2152. doi: 10.1177/0956797613488145
- Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). “ELAN: a professional framework for multimodality research,” in *5th International Conference on Language Resources and Evaluation (LREC 2006)*, 1556–1559.
- Xu, D., Yapanel, U., and Gray, S. (2009). *Reliability of the LENA™ Language Environment Analysis System in Young Children’s Natural Home Environment*. Available online at: https://www.lena.org/wp-content/uploads/2016/07/LTR-05-2_Reliability.pdf (accessed July 28, 2021).
- Xu, T. L., de Barbaro, K., Abney, D. H., and Cox, R. (2020). Finding structure in time: visualizing and analyzing behavioral time series. *Front. Psychol.* 11:1457. doi: 10.3389/fpsyg.2020.01457
- Young, S., and Gillen, J. (2007). Toward a revised understanding of young children’s musical activities: reflections from the “Day in the Life” project. *Curr. Musicol.* 84, 7–27.
- Zipf, G. K. (1936). *The Psychobiology of Language: An Introduction to Dynamic Philology*. Oxford: Routledge.
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley Press.
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Mendoza and Fausey. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Global Perspective on Testing Infants Online: Introducing ManyBabies-AtHome

Lorijn Zaadnoordijk^{1*}, Helen Buckler², Rhodri Cusack¹, Sho Tsuji³ and Christina Bergmann⁴

¹Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland, ²School of English, University of Nottingham, Nottingham, United Kingdom, ³International Research Center for Neurointelligence, The University of Tokyo, Tokyo, Japan, ⁴Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

OPEN ACCESS

Edited by:

Klaus Libertus,
University of Pittsburgh,
United States

Reviewed by:

Minxuan He,
University of Maryland,
College Park, United States
Gabriela Markova,
University of Vienna, Austria

*Correspondence:

Lorijn Zaadnoordijk
l.zaadnoordijk@tcd.ie

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 April 2021

Accepted: 09 August 2021

Published: 09 September 2021

Citation:

Zaadnoordijk L, Buckler H, Cusack R,
Tsuji S and Bergmann C (2021) A
Global Perspective on Testing Infants
Online: Introducing
ManyBabies-AtHome.
Front. Psychol. 12:703234.
doi: 10.3389/fpsyg.2021.703234

Online testing holds great promise for infant scientists. It could increase participant diversity, improve reproducibility and collaborative possibilities, and reduce costs for researchers and participants. However, despite the rise of platforms and participant databases, little work has been done to overcome the challenges of making this approach available to researchers across the world. In this paper, we elaborate on the benefits of online infant testing from a global perspective and identify challenges for the international community that have been outside of the scope of previous literature. Furthermore, we introduce ManyBabies-AtHome, an international, multi-lab collaboration that is actively working to facilitate practical and technical aspects of online testing and address ethical concerns regarding data storage and protection, and cross-cultural variation. The ultimate goal of this collaboration is to improve the method of testing infants online and make it globally available.

Keywords: global collaboration, replicability, method development, online testing, infancy

INTRODUCTION

Online testing holds vast promise for infant scientists. Conducting developmental research online can foster innovation, impact, and access (e.g., Sheskin et al., 2020) by allowing access to larger, more diverse samples and by creating cost-efficient joint participant databases for easier recruitment. Such possibilities facilitate more reproducible science and ultimately create opportunities to investigate questions that are uniquely accessible by testing diverse populations of infants in their natural home environment and/or in large samples.

The past years have seen a lot of advances on this front: Platforms designed specifically for developmental research (e.g., Scott and Schulz, 2017; Lo et al., 2021a), as well as language-specific participant recruitment initiatives (e.g., ChildrenHelpingScience.com; KinderSchaffenWissen.de) are being developed, and online studies are being conducted (e.g., Scott et al., 2017; Tran et al., 2017; Rhodes et al., 2020).

Although these initiatives provide a useful basis for creating infrastructures for online testing, efforts to overcome language, cultural, and regulatory barriers are still scarce: Practical recommendations and software solutions tend to assume US-based research or, in rare cases, are initiatives within the confines of another country or region. In addition, they do not

address all needs of the developmental science community, such as the feasibility of paradigms for online testing (see Scott and Schulz, 2017 for an exception).

In this paper, we aim to provide a global perspective on online infant testing, identifying challenges for the international community that were outside of the scope of previous literature. We then introduce ManyBabies-AtHome, an international, multi-lab effort to improve methods of testing infants online. This collaboration of labs distributed across all populated continents is actively working to facilitate practical and technical aspects of online testing as well as address ethical concerns regarding data storage, data protection, and cross-cultural differences. First, however, we will describe the motivations behind testing infants online.

THE “WHAT” AND “WHY” OF TESTING INFANTS ONLINE

For acquiring data online, several options are available. Apps and games can be used to administer parental questionnaires (e.g., Mayor and Mani, 2019; Chai et al., 2020) or acquire child data (e.g., Frank et al., 2016; Semmelmann et al., 2016; Lo et al., 2021b). Researchers can also conduct experiments while in a video call with the participants (i.e., synchronous testing). This method requires coordination between parent and researcher and imposes the schedule of the researcher as a limiting factor. Therefore, many researchers have turned to asynchronous, browser-based testing, the focal method in this article. In asynchronous testing, parents and their infants participate in experiments at a time that is convenient to them, without an experimenter present. Relevant information (e.g., the infant's date of birth) is logged and a webcam recording may be made of the infant doing a task on the computer (e.g., looking or touching), with the parent (e.g., reading a book or playing together), or away from the computer (e.g., playing with toys and vocalizing). The data are sent to the experimenter, who can review them at their own time. Asynchronous online testing has several benefits compared to lab-based testing.

The first benefit, participant sample, is 2-fold and pertains both to sample size and sample diversity. Many studies in developmental psychology suffer from low statistical power (Bergmann et al., 2018) due to small sample sizes and limited number of observations per participant (Byers-Heinlein et al., 2021). Online testing has the potential to allow researchers to test larger samples in less time, because (1) participants can participate in parallel; (2) there is no need to schedule the session; and (3) the study is accessible to participants who cannot come to the lab. The latter also means access to more diverse samples (Scott and Schulz, 2017; Rhodes et al., 2020; Cuccolo et al., 2021), such as people who do not live close to research labs or who work full time but who do have access to a computer with an Internet connection. This is important within a country as well as globally. Approximately 12% of the global population is western, educated, industrialized, rich, and democratic (WEIRD), but they make up 80% of

participants in psychology experiments (Henrich et al., 2010; see also Nielsen et al., 2017). Conclusions based on these participants may not generalize to the remaining 88% of the population. Online testing, thus, has the potential to improve the robustness of our studies due to well-powered studies, to increase the representativeness of the sample to match the global population, and to increase the ease of testing the generalizability of one's findings across various demographics.

A second benefit is increased replicability of the experimental protocol. Codified and fully automatized online experiments are easily replicable and extendable: All details related to the design, protocol, instructions, and testing session are specified in sharable and reusable code, materials, and text. This also facilitates collaborations between labs across the world as everyone can use the same protocol and there is no need for specific lab equipment.

A third benefit is reduced cost for the researcher. Running studies online is less labor-intensive and thus cheaper. Especially when testing asynchronously, there is a substantial reduction in the number of hours spent on scheduling and lab visits. Online testing is quick; researchers can in principle recruit and test hundreds of participants in 1 day (Berinsky et al., 2012; Casler et al., 2013). Finally, studies can be done at the infant's convenience, potentially increasing the chance of successful data acquisition, leading to fewer dropouts – whose data acquisition cost time and who may still receive rewards.

The main benefits of online testing thus can be summarized as increased size and diversity of the sample, more replicable and extendable experiments, easier collaboration, and lower cost. The recent pandemic has emphasized an additional benefit: avoiding the risk of infection. This is worth considering more generally, especially when working with physically vulnerable populations, such as infants. The accessibility of the method may also have benefits in clinical settings, for instance for developmental follow-ups. Because of all these benefits, we predict sustained interest in and use of online methods.

CHALLENGES ASSOCIATED WITH TESTING INFANT ONLINE

Testing infants online also comes with challenges, and many of which are additionally problematic for a global perspective and international collaborations. Software solutions are often inaccessible to large parts of the world due to being optimized for a certain country, law, culture, and/or language. Since current solutions cover North America and some of Europe, the WEIRD bias in participant sampling may be reinforced. It is outside the scope of this paper to discuss all challenges in detail, but in this section, we aim to raise general awareness about the current limits of broadly adopting online testing.

Laws and regulations form the first challenge. Local laws and regulations vastly differ regarding data collection, storage, and sharing. Using US-based platforms might be a problem under the European General Data Protection Regulation (GDPR), for example, because of concerns about who has access to data stored outside of Europe. The vague language and non-static

nature of the regulations (see, for instance, the United Kingdom following Brexit) and variability in local interpretations (e.g., Clarke et al., 2019) mean that researchers often are not aware of their options (Greene et al., 2019). This means that Research Ethics Committees (RECs) make decisions based on individual interpretations, causing an additional source of variability. As predicted by Litton (2017), even RECs that are governed under the same law can disagree on consent forms, use of US-based corporate cloud services, reimbursements, and so on. Although this is not specific to online testing, the novelty of the method and technology involved means there is no commonly accepted standard yet, causing a greater degree of unpredictability regarding REC decisions. This makes it difficult to make general recommendations or exchange experiences.

A second challenge pertains to international and cross-cultural data acquisition. Most platforms have been developed in one language (often English). This limits the possibility for global data collection. In addition to the language *per se*, which could be resolved with a translation, there are important cultural and contextual differences that need to be considered, such as a conversion between educational degrees, the formality of language use, and culturally sensitive approach to topics like asking about health and developmental delays. This means that all materials – from landing page to questionnaires – must not only be translated but also be culturally adapted (see Beaton et al., 2002).

A third challenge concerns the accessibility of online testing. Although online testing offers great potential for acquiring larger and more diverse samples, it is important to realize its limitations in terms of accessibility. Online testing relies on access to the Internet, not just for the experiment itself but often also for participant recruitment. Some populations will be easier to recruit *via* Internet advertising and social media presence than others. The best ways of recruiting various subpopulations for online infant testing have yet to be systematically investigated. Moreover, in online testing, the experiment and data quality are determined by participants' equipment and Internet connection at home. Researchers must consider the study's equipment and technical requirements, as these may limit data acquisition in certain subpopulations or countries. Online testing has the potential to reach more people but is not yet able to reach everyone. Fortunately, computers, webcams, and Internet connections are becoming increasingly accessible with nearly 50% of the world population using the Internet in 2017, and 16.3% of individuals ranked as having a low income using the Internet in 2017 compared to 2.2% just 10 years earlier.¹

A fourth challenge is obtaining high-quality data. While this challenge is not unique to the issue of globalization, its resolution requires a broad, collaborative perspective. Compared to a lab setting, online testing means less control over factors commonly associated with data quality in infant research. Precise temporal measures and reliance on exact timings can pose challenges (Anwyl-Irvine et al., 2020; Bridges et al., 2020). Furthermore, the parent implicitly takes on the role as co-experimenter regarding, for example, the lighting conditions

for any type of video recording, infant positioning, and the presence of distractions. This role of the parent as co-experimenter increases the need for clear and appropriate instructions to ensure good data quality. Furthermore, it may be necessary to expect a higher attrition rate for online studies than lab-based studies due to problems with data quality. Acquiring high-quality data are also a critical prerequisite for automatic coding of participants' behavior, such as looking behavior from webcam recordings. The latter is still subpar to eye-trackers in the lab; even simply tracking whether an infant is looking to the screen is not accurate enough to be used for infant-controlled procedures (Chouinard et al., 2019). Finally, asynchronous online testing removes certain sources of variability (e.g., differences in protocols between labs), but it likely introduces other sources of noise (e.g., distractors in the environment, increased parental interference, and feasibility to develop a robust online procedure for certain research questions or paradigms). Larger sample sizes and clear parental instructions may counteract some of this noise, but this may not be a solution for all types of research questions. The limitations mentioned in this section should be taken into account when deciding whether to conduct the study in the lab or online.

INTRODUCING ManyBabies-AtHome

To bundle the field's knowledge and advance online testing of infants, a large-scale collaboration, the ManyBabies-AtHome (MBAH) project has been initiated. MBAH is an independent project within the ManyBabies consortium² that aims to contribute to best research practices and universally replicable studies in all sub-fields of developmental science (e.g., language development, learning mechanisms, and social cognition). While previous ManyBabies projects have focused on the validity and replicability of specific findings and theories (see, e.g., Frank et al., 2017; Byers-Heinlein et al., 2020; ManyBabies Consortium, 2020; Visser et al., 2021), the MBAH project focuses on collaborative methods development for global online infant testing.

MBAH advances online testing efforts by (1) assessing the community's needs and wishes, (2) establishing generally applicable solutions in procedure, documentation, and analysis to make online testing methods accessible and robust across the world and to provide templates and materials for reuse and adaptation, (3) conducting studies to develop and test various paradigms for their suitability and robustness in the context of online infant testing, and (4) collecting and annotating a large dataset of infant gaze data that can be exploited for the development of automatic gaze coding approaches, which are necessary for infant-controlled paradigms.

Assessing the Needs of the Community

To understand the needs of the community, we conducted two informal surveys of researchers engaged with MBAH in

¹data.worldbank.org

²<https://manybabies.github.io/projects/>

spring 2020.³ In the first survey, we asked what methods or paradigms they use in the lab and what methods they would like to use for online testing. The responses indicate that researchers in our consortium are looking for the online version of a variety of paradigms, from preferential looking to parent-child interactions. We further asked the consortium which method or paradigm they would prioritize, which led to a strong support for looking behavior studies and preferential looking in particular. The second survey focused on ethics, data protection, and laws and regulations. At the time, many members of the consortium were subject to GDPR, which had come into effect 2 years prior. It is noteworthy that many researchers did not know, at the time of the survey, which options for reimbursements, data storage, and software solutions would be accepted by their RECs or local laws.

Procedures and Methods

New methods give rise to new questions about procedures. From ethics applications to data analysis, the research community needs to explore the possibilities and will ideally agree on acceptable standards. Direct collaboration within the community *via* MBAH makes this process more efficient and allows researchers to immediately voice their concerns and opinions. This allows us to take cross-cultural considerations, language barriers, and local laws and regulations into account. In collaboration with the consortium, we are making sure that data acquisition and storage meet their local requirements (see also Section “Ethics and Data Protection”). Furthermore, in addition to translating all Web sites visible to the parents and adapting the language use to cultural norms, we work together as a team to make sure the selected stimuli are appropriate and meaningful across cultures and languages (see also Section “Cultural Barriers”). MBAH is thus able to explore the possibilities and evaluate the benefits and downsides of various aspects of online infant testing across the world. Moreover, individual researchers will benefit from the knowledge acquired through MBAH regarding, for instance, the write-up of ethics applications, recruitment of participants, and instructions for parents.

MBAH Studies

Studies within MBAH are grassroots efforts, where members of the community can propose paradigms to study a research question on any aspect of development and if there is sufficient interest, efforts for joint study design are pooled. The first MBAH studies are efforts designed to suit the unique context of developing online testing methods. MBAH's initial focus is on studies using looking behavior as the primary measure. We, the MBAH steering committee, opted for asynchronous testing because of its benefits (see Section “The “What” and “Why” of Testing Infants Online”). We further decided, after reviewing several options in summer 2020, to conduct our studies on the LookIt platform (Scott and Schulz, 2017), as

this platform is designed specifically for asynchronous testing of infant looking behavior studies and is well tested, supported, and documented. This decision poses certain challenges as it is a US-based platform that uses the commercial cloud for data storage. However, if groups across the world are to be enabled and encouraged to acquire globally representative samples, it will be essential to break down data silos, wherever they may be. We are therefore focusing our efforts on ensuring only de-identified data are shared, while explaining this process to participants, and describing the scientific case for international data sharing to RECs. We also welcome parallel data collection using different platforms, such as Gorilla, but focus our efforts on supporting the use of LookIt. Our hope is that developed materials can be used across platforms.

Study 1: Proof-of-Concept Study

This study's primary goal is to work out general issues of online testing with an international consortium, including practical matters relating to ethics, data protection, and translation/cultural adaptation. As a secondary goal, this first study uses a preferential looking paradigm to assess infants' preference for static vs. moving images. This preference has been established in the lab (e.g., Shaddy and Colombo, 2004) and therefore makes for a good proof-of-concept study. With this paradigm, we can further assess previous general findings relating to infant looking time, such as whether infants' looking time decreases with age (Colombo, 2001; Courage et al., 2006). MBAH, like all ManyBabies projects, is committed to transparent and open science and will pre-register the hypotheses and analysis plan for this study.

Planned Future Studies

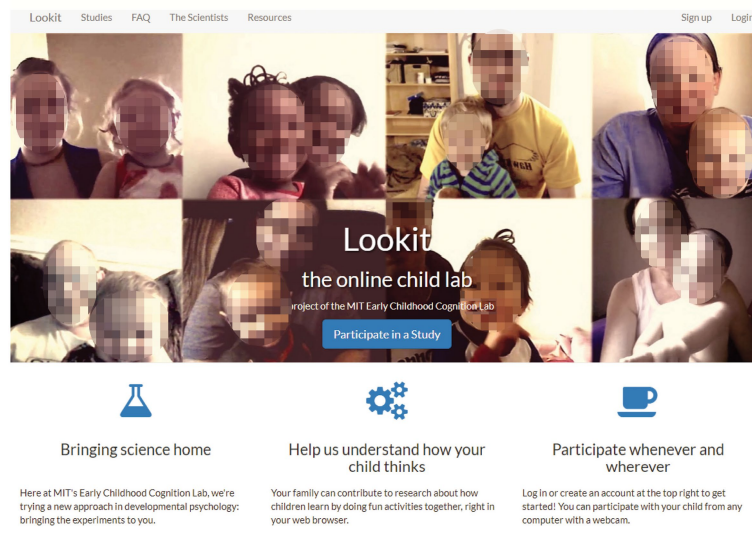
After evaluating whether the procedures developed for the proof-of-concept study are feasible for the developmental science community globally, we will conduct several studies. In part, these will be expanding on Study 1 by addressing other research questions with preferential looking paradigms. Furthermore, MBAH will increase its range of paradigms to include, for instance, a looking-while-listening paradigm that is currently being developed and plans to move toward replications of lab studies. Due to the broad range of expertise within the consortium, the feasibility of these paradigms can be assessed and adjusted at each step in the process. We thus follow the ManyBabies tradition of working toward a consensus-based best test of a phenomenon.

Automatic Gaze Coding

The data acquired across MBAH studies will be pooled in a rich, annotated dataset. This dataset will serve as the basis for developing and improving automatic gaze coding algorithms. Although webcam-based gaze tracking for adults has made considerable progress (Sammelmann and Weigelt, 2018), tracking infants' looks are still challenging (Chouinard et al., 2019), causing most labs to resort to manual coding. However, in addition to being labor-intensive, manual coding introduces inter- and intra-rater variability, leading to additional noise,

³Since these were informal online questionnaires without informed consent procedures, we cannot publish the actual data and will instead present a qualitative summary of researchers' responses.

Original LookIt Homepage (in English)



Translated LookIt Homepage (in Japanese)



FIGURE 1 | The original LookIt homepage (<http://lookit.mit.edu>) and an example of one of the translations (Japanese) currently in progress.

which could be prevented with reliable automated methods. The algorithms will initially be developed for post-hoc offline gaze tracking. However, our goal is to develop online gaze-tracking algorithms, which would allow for infant-controlled paradigms, such as habituation studies.

Current Challenges for MBAH Ethics and Data Protection

The variability in local laws and the lack of knowledge among researchers regarding the tools and processes that are available to them poses a challenge to composing ethics and data protection protocols that will be acceptable for the RECs of all our consortium members. The main complications relate to data storage and data sharing. We are working on solutions

for researchers whose local regulations limit global data storage and sharing (e.g., those based in the EU) to enable them to acquire data too. We aim to obtain umbrella approval for MBAH, which should allow most consortium members to acquire data. Researchers may also apply with their local ethics boards if they need to meet specific criteria that our umbrella approval does not cover. We are committed to finding solutions for all researchers in our consortium.

Cultural Barriers

Since LookIt, the main platform that will be used for MBAH, is currently targeted at the English-speaking population in the United States, our first objective is to translate this platform into other languages. However, adaptation to other languages

and locations goes beyond translation and requires continuous checks for suitability of all questions. In some countries, it is impolite or even illegal to solicit information about infant health (e.g., developmental delays). On the technical side, this requires ensuring that all parts of the platform are contained in files that can be subject to language selection. The translation of the general LookIt pages (such as the homepage, the user's profile page, and the FAQ) as well as the MBAH study-specific pages of 16 languages is currently in progress (see **Figure 1**), and MBAH welcomes anyone who speaks English and another language and wants to contribute to our translating effort to join the project. Our translations will hopefully make LookIt a more viable choice for individual researchers outside of the MBAH project as well.

Contributing to MBAH

MBAH welcomes all researchers and other interested parties to contribute and aims to create an inclusive and diverse environment. Contributors may be at any stage in their career (student to professor), may be from any country, do not have to have participated in earlier ManyBabies projects, do not need to be members of any society, and do not need to know the leadership team to be involved. Interest in potential contribution is free of commitment and can be expressed by email to any of the members of the leadership team (i.e., the authors of this paper), who will send the relevant information. We are keeping track of various types of contributions (according to CRediT principles), which result in authorship on corresponding project papers. For secondary analyses, we aim to openly share anonymized data summaries and where possible (depending on ethical approval and parental consent) share the raw video data.

MBAH plans to also incorporate existing platforms and procedures for data acquisition, processing (e.g., annotation and anonymization), storage, and management. For data acquisition, we have focused on LookIt as our main platform. We recognize, however, that some ethics boards might not approve the use of this US-based platform, in which case data may be acquired elsewhere too. We welcome solutions for any of the above-mentioned data-related processes from research groups, platforms, and companies.

CONCLUSION

Online testing offers great potential as a new tool in the developmental scientist's toolbox as it increases participant diversity, replicability, transparency, and collaborative possibilities and reduces costs for researchers and participants. It has benefits beyond scientific practice too, as it may increase the possibilities for and accessibility of clinical developmental follow-ups. However, despite the rise of platforms and participant databases

that make online testing possible, little work has been done to overcome the challenges of making this approach globally available.

Here, we have introduced the international, multi-lab MBAH project. MBAH works to address and resolve the challenges and to create generally applicable solutions in procedure, documentation, and analysis to make online infant testing methods accessible and robust across a range of home environments across the world. Hurdles that are revealed will be resolved in a community-based manner, allowing for rapid and direct input from researchers from different countries and cultures. To accomplish this, we welcome researchers at all levels to join our consortium.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The authors of this article are the full leadership team of MBAH. As such, LZ, HB, RC, ST, and CB have contributed (and are contributing) to the conceptualization, implementation, creating documentation, and funding acquisition related to the MBAH project. LZ wrote the first draft and revision of this article. All authors contributed to the article and approved the submitted version.

FUNDING

MBAH is funded by a JST ACT-X grant in the research area "AI powered Research Innovation/Creation" awarded to ST. Moreover, MBAH is facilitated by grants awarded to individual authors: MSCA Individual Fellowship (InterPlay, #891535) awarded to LZ, the ERC Advanced grant (FOUNDCOG, #787981) awarded to RC, and the Jacobs Foundation Fellowship awarded to ST.

ACKNOWLEDGMENTS

We would like to thank the ManyBabies Governing Board, the ManyBabies-AtHome translating team (in particular Ana Maria Portugal), and the ManyBabies-AtHome consortium for their support and their contributions to the project. We thank Kim Scott and the Lookit team for their collaboration with ManyBabies-AtHome. Finally, we would further like to thank the reviewers for their constructive feedback.

REFERENCES

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., and Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behav. Res. Methods* doi: 10.3758/s13428-020-01501-5 [Epub ahead of print].

Beaton, D. E., Bombardier, C., Guillemin, E., and Ferraz, M. B. (2002). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 25, 3186–3191. doi: 10.1097/00007632-200012150-00014

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., et al. (2018). Promoting replicability in developmental research through

- meta-analyses: insights from language acquisition research. *Child Dev.* 89, 1996–2009. doi: 10.1111/cdev.13079
- Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Polit. Anal.* 20, 351–368. doi: 10.1093/pan/mpr057
- Bridges, D., Pitiot, A., MacAskill, M. R., and Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ* 8:e9414. doi: 10.7717/peerj.9414
- Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., et al. (2020). Building a collaborative psychological science: lessons learned from ManyBabies 1. *Can. Psychol.* 61, 349–363. doi: 10.1037/cap0000216
- Byers-Heinlein, K., Bergmann, C., and Savalei, V. (2021). Six solutions for more reliable infant research. [Preprint].
- Casler, K., Bickel, L., and Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Hum. Behav.* 29, 2156–2160. doi: 10.1016/j.chb.2013.05.009
- Chai, J. H., Lo, C. H., and Mayor, J. (2020). A Bayesian-inspired item response theory-based framework to produce very short versions of MacArthur-bates communicative development inventories. *J. Speech Lang. Hear. Res.* 63, 3488–3500. doi: 10.1044/2020_JSLHR-20-00361
- Chouinard, B., Scott, K., and Cusack, R. (2019). Using automatic face analysis to score infant behaviour from video collected online. *Infant Behav. Dev.* 54, 1–12. doi: 10.1016/j.infbeh.2018.11.004
- Clarke, N., Vale, G., Reeves, E. P., Kirwan, M., Smith, D., Farrell, M., et al. (2019). GDPR: an impediment to research? *Ir. J. Med. Sci.* 188, 1129–1135. doi: 10.1007/s11845-019-01980-2
- Colombo, J. (2001). The development of visual attention in infancy. *Annu. Rev. Psychol.* 52, 337–367. doi: 10.1146/annurev.psych.52.1.337
- Courage, M. L., Reynolds, G. D., and Richards, J. E. (2006). Infants' attention to patterned stimuli: developmental change from 3 to 12 months of age. *Child Dev.* 77, 680–695. doi: 10.1111/j.1467-8624.2006.00897.x
- Cuccolo, K., Irgens, M. S., Zlokovich, M. S., Grahe, J., and Edlund, J. E. (2021). What crowdsourcing can offer to cross-cultural psychological science. *Cross-Cult. Res.* 55, 3–28. doi: 10.1177/1069397120950628
- Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., et al. (2017). A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy* 22, 421–435. doi: 10.1111/infa.12182
- Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., and Yurovsky, D. (2016). Using tablets to collect data from young children. *J. Cogn. Dev.* 17, 1–17. doi: 10.1080/15248372.2015.1061528
- Greene, T., Shmueli, G., Ray, S., and Fell, J. (2019). Adjusting to the GDPR: The impact on data scientists and behavioral researchers. *Big Data* 7, 140–162. doi: 10.1089/big.2018.0176
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83. doi: 10.1017/S0140525X0999152X
- Litton, J. E. (2017). We must urgently clarify data-sharing rules. *Nature News* 541:437. doi: 10.1038/541437a
- Lo, C. H., Mani, N., Kartushina, N., Mayor, J., and Hermes, J. (2021a). e-Babylab: An open-source browser-based tool for unmoderated online developmental studies. *PsyArXiv* [Preprint].
- Lo, C. H., Rosslund, A., Chai, J. H., Mayor, J., and Kartushina, N. (2021b). Tablet assessment of word comprehension reveals coarse word representations in 18–20-month-old toddlers. *Infancy* 26, 596–616. doi: 10.1111/infa.12401
- ManyBabies Consortium (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* 3, 24–52. doi: 10.1177/2515245919900809
- Mayor, J., and Mani, N. (2019). A short version of the MacArthur-bates communicative development inventories with high validity. *Behav. Res. Methods* 51, 2248–2255. doi: 10.3758/s13428-018-1146-0
- Nielsen, M., Haun, D., Kärtner, J., and Legare, C. H. (2017). The persistent sampling bias in developmental psychology: a call to action. *J. Exp. Child Psychol.* 162, 31–38. doi: 10.1016/j.jecp.2017.04.017
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (part 2): assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/OPMI_a_00001
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Semmelmann, K., Nordt, M., Sommer, K., Röhne, R., Mount, L., Prüfer, H., et al. (2016). U can touch this: how tablets can be used to study cognitive development. *Front. Psychol.* 7:1021. doi: 10.3389/fpsyg.2016.01021
- Semmelmann, K., and Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: a first look. *Behav. Res. Methods* 50, 451–465. doi: 10.3758/s13428-017-0913-7
- Shaddy, D. J., and Colombo, J. (2004). Developmental changes in infant attention to dynamic and static stimuli. *Infancy* 5, 355–365. doi: 10.1207/s15327078in0503_6
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Tran, M., Cabral, L., Patel, R., and Cusack, R. (2017). Online recruitment and testing of infants with Mechanical Turk. *J. Exp. Child Psychol.* 156, 168–178. doi: 10.1016/j.jecp.2016.12.003
- Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S. H., et al. (2021). Improving the generalizability of infant psychological research: the ManyBabies model. *Behav. Brain Sci.* [Preprint]. doi: 10.31234/osf.io/8vwbw

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Zaadnoordijk, Buckler, Cusack, Tsuji and Bergmann. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



“May I Grab Your Attention?”: An Investigation Into Infants’ Visual Preferences for Handled Objects Using Lookit as an Online Platform for Data Collection

Christian M. Nelson and Lisa M. Oakes*

Department of Psychology and the Center for Mind and Brain, University of California, Davis, Davis, CA, United States

OPEN ACCESS

Edited by:

Eric A. Walle,
University of California, Merced,
United States

Reviewed by:

Eliza L. Nelson,
Florida International University,
United States
Melissa M. Kibbe,
Boston University, United States

*Correspondence:

Lisa M. Oakes
lmoakes@ucdavis.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 June 2021

Accepted: 12 August 2021

Published: 10 September 2021

Citation:

Nelson CM and Oakes LM (2021)
“May I Grab Your Attention?”: An
Investigation Into Infants’ Visual
Preferences for Handled Objects
Using Lookit as an Online Platform for
Data Collection.
Front. Psychol. 12:733218.
doi: 10.3389/fpsyg.2021.733218

We examined the relation between 4- to 12-month-old infants’ ($N = 107$) motor development and visual preference for handled or non-handled objects, using Lookit (lookit.mit.edu) as an online tool for data collection. Infants viewed eight pairs of objects, and their looking was recorded using their own webcam. Each pair contained one item with an easily graspable “handle-like” region and one without. Infants’ duration of looking at each item was coded from the recordings, allowing us to evaluate their preference for the handled item. In addition, parents reported on their infants’ motor behavior in the previous week. Overall, infants looked longer to handled items than non-handled items. Additionally, by examining the duration of infants’ individual looks, we show that differences in infants’ interest in the handled items varied both by infants’ motor level and across the course of the 8-s trials. These findings confirm infant visual preferences can be successfully measured using Lookit and that motor development is related to infants’ visual preferences for items with a graspable, handle-like region. The relative roles of age and motor development are discussed.

Keywords: infant, visual preference, motor development, online testing, Lookit

As infants achieve motor milestones, they gain access to new information about the objects around them. Infants who sit up can pick up objects and look at them from many angles, and infants who crawl can see and move to objects in the distance. Thus, changes in infants’ bodies and motor abilities help determine what information they have access to, attend to, and learn about (Kretch et al., 2014; Smith et al., 2018). For example, the emergence of sitting and changes in object manipulation are associated with infants’ 3-D object completion abilities (Soska et al., 2010), attention to object features in dynamic events (Perone et al., 2008; Baumgartner and Oakes, 2013), and figure-ground segregation (Ross-Sheehy et al., 2016). Moreover, experience reaching and grasping objects with “sticky mittens” can induce changes in infants’ attention to objects (Needham et al., 2002), interest in faces (Libertus and Needham, 2011), and perception of causal interactions (Rakison and Krogh, 2012). Shifts from crawling to standing and walking are associated with mental rotation ability (Frick and Möhring, 2013) and changes in how infants initiate eye contact with caregivers (Yamamoto et al., 2019). The emergence of walking comes with even more variation in experience with distal objects in the environment (Karasik et al., 2011). Taken together, it is clear that motor achievements in the first year have cascading effects on other aspects of development (Oakes and Rakison, 2019).

In the present investigation, we ask how infants' visual preference for items with a more easily graspable region is related to changes in their motor abilities. We focused on potentially graspable objects because Corbetta et al. (2018) describe a perception-action loop by which infants reach for objects they see, inducing changes in their reaching, manual exploration, and visual inspection of those objects. Additionally, Libertus et al. (2013) found that infants with more reaching experience shift from an initial preference for larger, more salient objects toward studying the features of smaller, more graspable objects. In the current experiment, we included a wide range of developmental achievements—examining the preference for potentially graspable objects in a sample spanning pre-sitting to walking infants—to capture changes in visual perception from the cascading effect of motor development across infancy (Oakes and Rakison, 2019; Iverson, 2021).

A secondary goal of this study was to demonstrate the effectiveness of answering such questions using methods designed to assess infant cognition while physically distant. The COVID-19 pandemic introduced unique challenges to studying infant development while physically distant. Methods of examining visual preferences at a distance may provide opportunities to study infant development even beyond the pandemic. The adoption of such tools will diversify developmental science by removing barriers to participate that have been a limitation of traditional methods.

We used Lookit, which was developed to assess visual preference using participants' own computer, monitor, and web camera (Scott and Schulz, 2017). Our goal was to examine infants' visual preference for handled objects over non-handled objects. In addition, we asked parents to report on their infants' sitting, crawling, standing, and walking in the previous week, so we could determine whether infants' preferences for handled objects in our task was related to motor development. We chose these milestones because: (1) Sitting has been associated with increases in exploratory behaviors (Soska and Adolph, 2014), better prehensile hand use (Rochat and Goubet, 1995), and looking preferences for graspable objects (Libertus et al., 2013); (2) Crawling experience has been associated with infant visual perception of objects (Cicchino and Rakison, 2008; Schwarzer et al., 2013; Gerhard and Schwarzer, 2018); and (3) Standing and walking are associated with changes in visual input (Libertus and Hauf, 2017; Franchak, 2018) and visual perception (Frick and Möhring, 2013).

METHOD

Participants

To be eligible, infants must be born at term and residing in the US. We collected data between 08/26/2020 and 01/26/2021, until we had recorded 159 sessions, anticipating a moderate effect size (e.g., ~ 0.5 Cohen's d) and that we may be unable to use half of the data collected.

We excluded 30 sessions because the infants were ineligible; the infant did not reside within the United States ($n = 6$), was premature ($n = 13$), participated multiple times ($n = 7$), or was outside our target age range ($n = 4$). We excluded an additional

24 sessions because of technical problems (e.g., no video data, slow upload speeds, $n = 15$), other problems or distractions (e.g., parent peeking throughout the session, infant's eyes not visible, $n = 6$), or lack of infant interest (see data processing below, $n = 1$). Our final sample was 107 infants (M age = 248.30 days, $SD = 70.26$, 39 girls; histogram of age distribution is in **Supplemental Materials**).

Infants were recruited via the Lookit recruiter (i.e., emails were sent to families with accounts in Lookit), social media (ads on Facebook), and emails to families who had expressed interest in participation (see Oakes et al., 2021 for details regarding identification and recruitment of infant participants). All families residing in the US received a \$5 Amazon gift card¹.

Our sample was racially diverse and highly educated. Of our 107 infants, 57 were White, one was Black/African American, 10 were Asian American, 38 were multiracial, and one was unreported. Regardless of race, 19 infants were Hispanic. One (or both) parents had at least some college in 103 of the families, neither parent had any college in three families, and one family declined to state parental education. Ninety-eight families reported income; 58 reported income over \$100,000, 29 reported income between \$50,000 and \$100,000, and 11 reported income less than \$50,000.

Stimuli

Stimuli were photographs of 16 real, unfamiliar objects selected from the NOUN database (Horst and Hout, 2016), eight with handle-like protrusions (see **Figure 1**). Because the objects were novel and the handle-like protrusions varied, any preference for the handled objects would be the result of the infants' perception of the difference between the two types of objects, and not related to their knowledge of or experience with those items.

Because the Lookit platform involves each family using their own computer and monitor for the test, precise measurements of the stimuli as they were shown to each infant are not possible. However, each stimulus occupied $\sim 5\%$ of the total display, and were separated by a distance that was $\sim 50\%$ of the display. Comparison of the proportion covered by handled ($M = 5.23\%$, $SD = 2.78\%$) and non-handled ($M = 4.72\%$, $SD = 1.78\%$) objects revealed that, on average, there were no significant differences in the sizes of the two sets of objects, $t_{(14)} = 0.44$.

The objects differed on many dimensions (e.g., color, presence of pattern). To ensure that the objects did not differ in physical salience, we calculated the physical salience of each object in each stimulus pair using the Graph Based Visual Salience toolbox (GBVS, Harel et al., 2007), using the default settings. We used the GBVS toolbox because other research suggested it best predicts infant looking (Pomaranski et al., in press). We averaged the salience for the pixels in areas of interest—a region in the display that contained the objects on each stimulus. Thus, our salience values reflect the average salience of these regions for each of the pairs in our stimuli. Comparison of the average salience level

¹IRB and funding disallowed participation by or compensation for non-US participants.

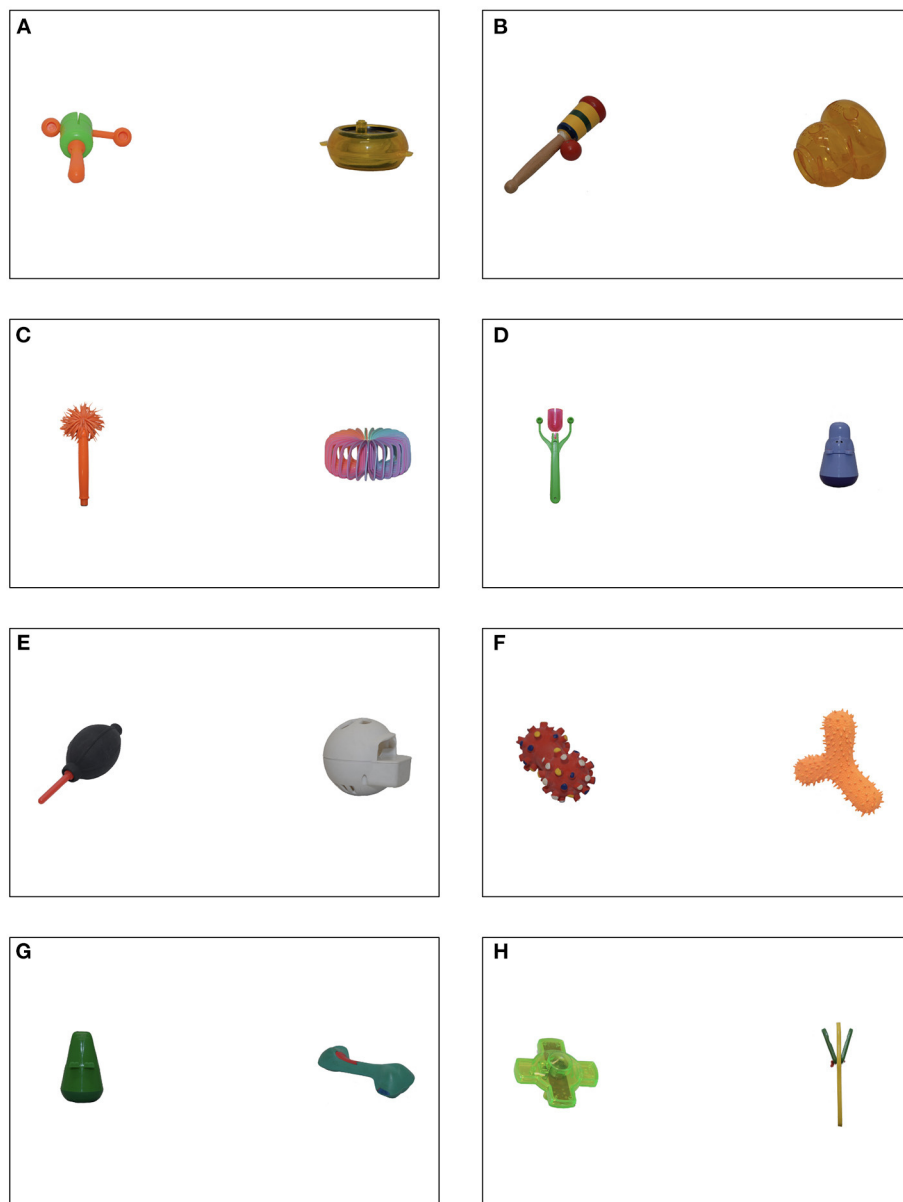


FIGURE 1 | The pairs of objects used in the study. Pairs (A) through (E) show the handled object on the left. Pairs (F) through (H) show the handled object on the right. Objects came from Horst and Hout (2016).

for handled and non-handled objects revealed no difference in salience, $t_{(14)} = 1.46$, $p = 0.17$.

Procedure

All sessions were conducted online, using Lookit. Parents created an account on Lookit, at which time they provide demographic information (e.g., state of residence, infant race, parent education). When ready to participate, parents logged into Lookit and selected our study. First, parents watched a short video describing our study. Then they read the consent document and verbally consented via video recording, as required on the

Lookit platform. Next, parents reported whether their infant had exhibited five different poses or behaviors in the previous week; each question was accompanied by images taken from the Alberta Infant Motor Scale (Piper et al., 1992) to depict motor milestones. Parents answered, “Yes”, “No” or “Unsure” regarding whether in the past week their infant (1) sat, (2) crawled (on belly or on hands and knees), (3) pulled to stand or (4) walked independently. For scoring, we determined the highest level that the parent said “yes” to (even if the previous levels were “no” or “unsure”), and classified infants as pre-sitters (score of 1, if parents said no or unsure to all of the behaviors), sitters (score

of 2, if parents said “yes” to sitting, but not to crawling, standing, or walking), crawlers (score of 3, if parents said “yes” to crawling, but not to standing or walking), standers (score of 4, if parents said “yes” to standing but not walking), or walkers (score of 5, if parents said “yes” to walking). Note that infants could “skip” a motor milestone (e.g., pull to stand or walk without crawling). We focused only on the highest motor milestone achieved, regardless of whether earlier milestones were skipped. Finally, parents viewed an instructional video illustrating how to hold their infant on their lap, facing the computer monitor (i.e., acting as a good “chair” for their infant), keeping their eyes closed during the session.

When ready, parents began the experimental session by pressing a key on their computer keyboard, which initiated a sequence of trials that continued without interruption (see **Figure 2**), unless the parent pressed the spacebar to pause (paused trials were excluded from the analyses). Each trial began with a 5-s attention-getting stimulus (i.e., a clip from an animated children’s movie or television show) presented at the center of the display. The experiment consisted of two trial blocks. The first trial of each block was a *calibration trial*, in which a looming object, accompanied by a jingling bell, appeared for 2.5-s first at the center, then to the left, then to the right, and finally to the center again. During these trials, at any given moment there was only one item present, directing infants’ attention to each location and allowing coders to calibrate their judgments about infant looks to the left or the right. After the calibration trials, there were four 8-s paired preference trials, each presenting a single pair of objects (one handled and one non-handled) accompanied by classical music. All infants saw the same eight pairs, which were divided into two blocks (A and B); within each block, the stimuli were presented in a random order for each infant. Twenty-seven infants received block A first followed by block B, and 80 infants received block B first. This uneven distribution resulted from us using one order for the first weeks of data collection and then switching the order. Because the trials are ordered randomly within blocks, this unbalanced design will have minimal impact on the results.

Coding

Trained undergraduate research assistants used Datavyu (<https://datavyu.org/>) to code infants’ looking on all trials. Because during the calibration trials there was only one stimulus presented at a time, the infants’ looking to the left or right on was less ambiguous than during the experimental trials when two stimuli were presented side-by-side. Coders used the calibration trials to provide clear examples of the particular infants’ left and right looks. Coders then viewed the paired preference trials; first they viewed the trials in real time and then on a second pass they used the Datavyu jog function to identify the start and end of each look to the left, right, or center (e.g., to the attention-getter at the start of the trial). A “look” consisted of at least three consecutive frames of gaze to the same location. The primary coder recorded looking on all trials, and these data were used in our analysis. A second reliability coder recorded the looking on two randomly selected test trials for each infant included in our final sample.

The frame-by-frame agreement between the two coders was, on average, 94% (range: 75% to 99%—103 were above 80%).

Data Processing

We calculated infants’ total looking to the handled and non-handled object on each trial (number of frames directed to each side X frame duration²) and the duration of each individual look, which was defined as any successive frames to the same region (left, right, or center). We calculated handle preference scores for the trial as a whole by dividing the infants’ looking to the handled object by their total looking to the two objects combined.

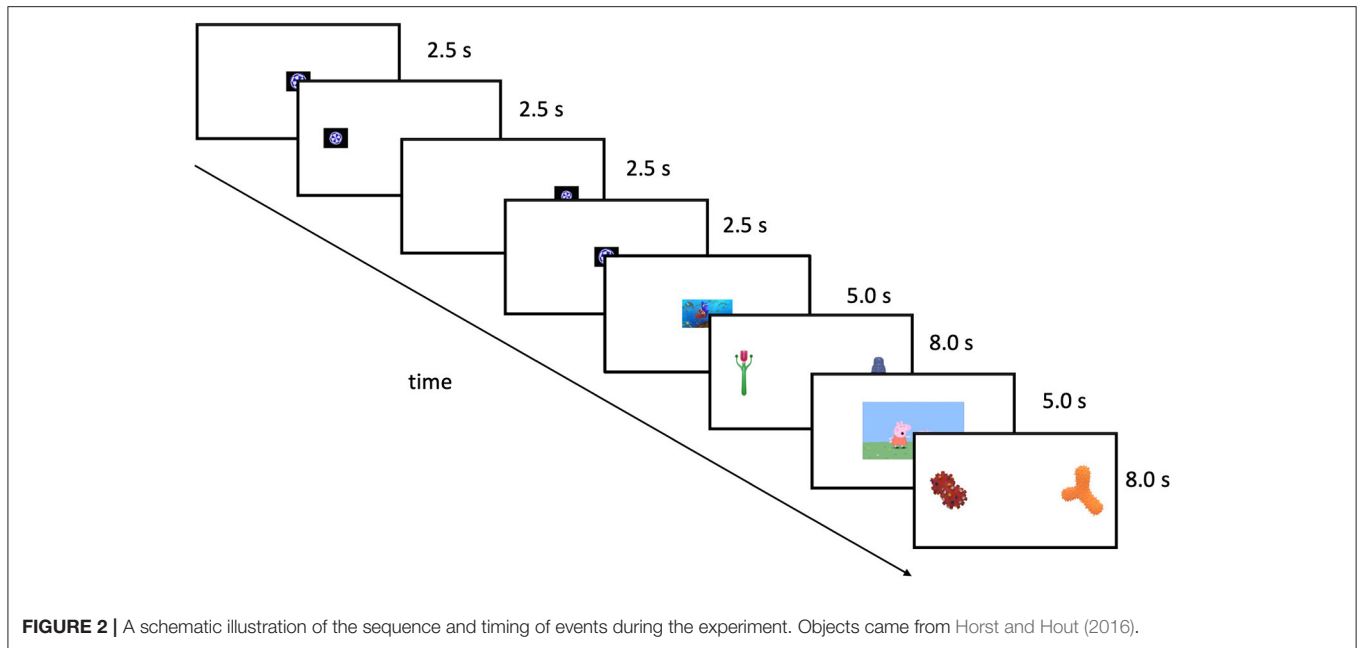
After participant exclusions, we evaluated 817 trials from 108 infants. We excluded 96 trials because the infant was fussy, failed to look, or their eyes were not visible; the parent looked at the experimental stimuli (a brief peek of no more than 1 s was allowed); or distraction (e.g. someone talked to and/or touched the infant, background noise). One infant was removed at this stage because they looked <2,000 ms on all trials. Thus, our analyses were conducted on 721 trials from our final sample of 107 infants.

Analysis Plan

All analyses were conducted in R (R Core Team, 2019). To provide an *overall* impression of the data, we calculated a single score for each infant by averaging their preference scores across trials. To understand how infants’ preference changed over time, we examined *trial-level* and *look-level* behavior with linear mixed-effects models, using the packages *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017). Omnibus *F*-statistics were used to evaluate the significance of the fixed effects from these models, and the *emmeans* package (Lenth et al., 2018) to extract marginal means from the omnibus *F*-test. These two models will allow us to determine how infants’ interest in the handled item relative to the non-handled item changed over time and trials, controlling for relative size and salience of the handled item on each trial. For each model, we first assessed the multicollinearity of the variables using the package *performance* (Lüdtke et al., 2021). Because we had no predictions related to infant sex, we did not include infant sex as a factor in our models. However, for transparency and consistency with NIH guidelines regarding reporting of sex as a biological factor, we disaggregate by sex when graphing our results.

To examine infants’ preference at the level of trial, we conducted an analysis with handle preference on each trial as the DV (handle preference was centered by subtracting chance, or 0.50, for ease of interpretation). We included fixed effects of motor level, trial number, and the interaction between these variables. We also included control variables of relative salience and relative size of the handled item. We included random effects of child and stimulus (i.e., the unique object pairs presented on each trial). An initial model revealed collinearity between age in days and motor level. Thus, our final model included only motor level, which was our variable of interest.

²Variation in framerates resulted from differences in web cameras, upload speed, etc. We used a custom script to extract the framerate from each video and determine the duration of each infants’ video frames.



Finally, we conducted a model with the duration of each individual look as DV. We included fixed effects of motor level, object type (handled or not), look index (e.g., whether it was the first look, second look, and so on), and interactions between these variables. We also included control variables of stimulus salience, and stimulus size, and random effects of child and stimulus. Again, an initial model revealed that age in days and motor level were collinear, so our final model included motor level.

RESULTS

Infants contributed on average 6.74 trials ($SD = 1.74$, range 1 to 8) to the analysis and looked on average 5079.3 ms ($SD = 1084.86$ ms) on each trial. Infants' age in days was not significantly correlated with average duration of looking, $r(107) = 0.004$, or the number of trials completed, $r(107) = 0.01$. Motor level also was not significantly correlated with average duration of looking, $rs^3(107) = 0.09$, or with the number of trials completed, $rs(107) = 0.11$. Unsurprisingly, motor score was significantly correlated with infants' age in days, $rs(107) = 0.87$, reflecting the fact that older infants were more motorically advanced than younger infants.

Our first analyses examined infants' overall handle preference, both averaged across all completed trials and examining preferences trial by trial. The average preference score for the group of infants as a whole was .54 ($SD = 0.10$), which was significantly greater than chance (0.50), $t_{(106)} = 4.23$, $p < 0.001$, $d = 0.41$. Planned comparisons conducted for each motor group revealed that only the locomotor infants (crawlers, standers, and walkers) had handle preferences that were significantly greater

than chance (see **Table 1**), however motor level and handle preference were not significantly related, $rs(107) = 0.12$.

We also conducted the LMM on infants' overall preference on each trial as specified earlier. This model revealed no significant effects of interactions. We conducted an analysis with age instead of motor level, which also did not reveal any significant effects or interactions (see **Supplementary Materials**).

Finally, we examined the duration of each individual look during a trial. Infants contributed, on average, 4.48 looks ($SD = 1.34$, range 1 to 12). On average looks to the handled objects were longer, $M = 1334$ ms, $SD = 502$, than to the non-handled objects, $M = 1119$ ms, $SD = 357$. Because the duration of looks that occur later in the trial are potentially constrained by the durations of earlier looks in that trial, we examined the distribution of look lengths and found that 75% of all looks were $< 1,500$ ms in duration. In addition, the duration of looks actually *increased* with increased index (see **Figure 3**). Thus, there is little evidence that long early looks suppressed the length of later looks.

We performed an LMM with duration of each individual look as the DV as described in the *Analysis Plan* section. This model showed a significant omnibus effects of look index, $F_{(1,3319.7)} = 57.75$, $p < 0.001$, due to infants' looks increased in duration across the trial, and salience, $F_{(1,1720.8)} = 13.42$, $p < 0.001$, due to infants' looks increasing with increased object salience.

Importantly, the model revealed a significant interaction between motor level and object type, $F_{(1,3244.8)} = 6.37$, $p = 0.012$, and a 3-way interaction with these two variables and look index, $F_{(1,3252.4)} = 4.17$, $p = 0.041$. Thus, motor level was related to the duration of infants' looks at handled versus non-handled objects. The 3-way interaction is displayed in **Figure 3**. Non-crawling infants (motor levels 1 and 2), demonstrated little or no difference in the duration of looks to handled and non-handled objects (with a suggestion that a difference may emerge in the later

³We conducted Spearman's Rank Order Correlations when examining relations with motor level, which was ordinal.

TABLE 1 | Mean handle preference and age (in days) by Motor Level.

| Motor Level | N | Mean Age | Mean handle preference | t | p | d | Scaled JZS Bayes Factor |
|-------------|----|-----------------------------|------------------------|------|------|------|-------------------------|
| Pre-sit | 15 | 149.67(sd = 26.14, 124–212) | 0.53 | 1.14 | 0.27 | 0.29 | BF ₀₁ = 2.19 |
| Sit | 18 | 205.94(sd = 41.88, 160–285) | 0.53 | 1.23 | 0.23 | 0.29 | BF ₀₁ = 2.14 |
| Crawl | 21 | 211.67(sd = 33.36, 124–259) | 0.55 | 2.38 | 0.03 | 0.52 | BF ₁₀ = 2.21 |
| Stand | 44 | 295.20(sd = 36.34, 206–362) | 0.54 | 2.12 | 0.04 | 0.32 | BF ₁₀ = 1.24 |
| Walk | 9 | 353.36(sd = 15.26, 328–375) | 0.61 | 3.74 | 0.01 | 1.25 | BF ₁₀ = 9.95 |

t-tests are one-sample t-tests comparing the means to chance (0.50). The Bayes Factors were calculated using the non-informative JZS prior with a scale factor of 0.707. BF₁₀ indicates support for the alternative hypothesis and BF₀₁ indicates support for the null hypothesis; BF between 1 and 3 provides anecdotal evidence, and BF between 3 and 10 provides moderate evidence.

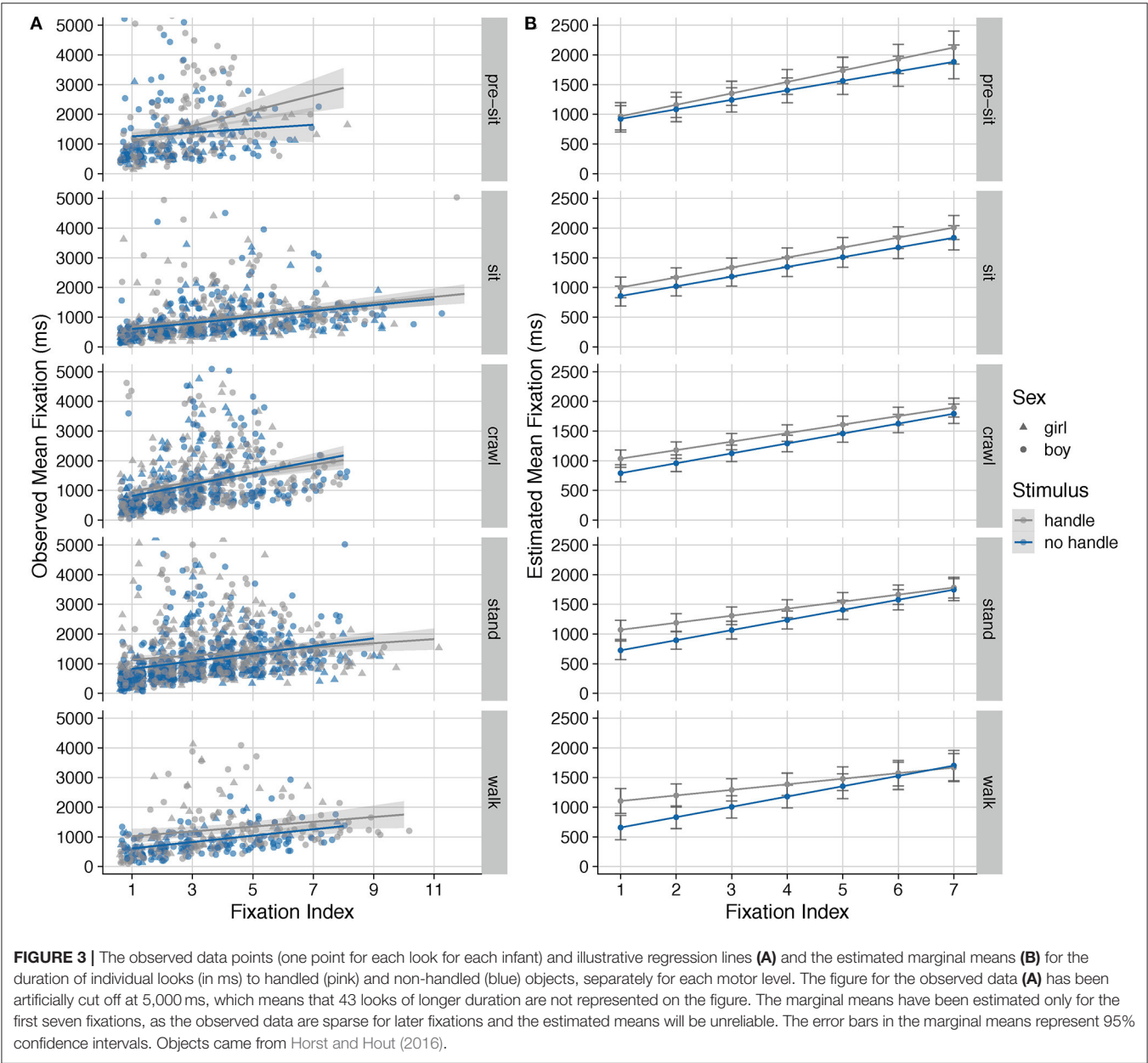


FIGURE 3 | The observed data points (one point for each look for each infant) and illustrative regression lines (A) and the estimated marginal means (B) for the duration of individual looks (in ms) to handled (pink) and non-handled (blue) objects, separately for each motor level. The figure for the observed data (A) has been artificially cut off at 5,000 ms, which means that 43 looks of longer duration are not represented on the figure. The marginal means have been estimated only for the first seven fixations, as the observed data are sparse for later fixations and the estimated means will be unreliable. The error bars in the marginal means represent 95% confidence intervals. Objects came from Horst and Hout (2016).

looks). For infants who are standing or walking (motor levels 4 or 5), early looks to the handled item are longer than early looks to the non-handled item. Thus, the three-way interaction stems from differences in the timing of when infants with different motor abilities show a preference for the handled item. Again, our analysis with age in days instead of motor development yielded no significant effects of or interactions with age (see **Supplemental materials**).

DISCUSSION

We observed that infants' visual object preferences were related to motor development, adding to a growing literature showing connections between motor development and infants' visual object perception and processing (Needham et al., 2002; Sommerville et al., 2005; Libertus and Needham, 2010; Soska et al., 2010). In addition, we successfully measured infants' visual preferences using online tools. These tools can be effective in advancing our understanding of infant cognitive development. Our findings confirm observations made by Scott et al. (2017) that meaningful data can be obtained using online tools, such as Lookit.

Our results are generally consistent with previous literature showing a relation between infants' object perception and motor development (Baumgartner and Oakes, 2013; Kretch et al., 2014; Franchak, 2018). Specifically, we observed that the duration of infants' individual looks to handled vs. non-handled items varied as a function of their motor level. Our results corroborate those of Libertus et al. (2013), who showed that infants with more reaching experience exhibited preferences for more graspable objects. Here we show that the cascading effect of multiple aspects of motor development influence how infants look at objects. Although the focus of previous research has been on reaching experience, achievements such as standing and walking also change infants' attention to, perception of, and interactions with objects (Karasik et al., 2011; Frick and Möhring, 2013). Thus, although our results cannot provide direct insight into why standing and walking would enhance infants' preference for handled objects *per se*, they are consistent with literature showing that object perception and preferences are related to gross motor development.

Of course, because motor level and age were confounded it is impossible to completely disambiguate them; it is possible that infants' increasing interest in handled objects is due to other factors (e.g., cortical maturation, experience). However, our findings suggest that changes in handle preference are due, at least in part, to motor development. First, because motor development increases with age, changes in motor abilities—and the interactions with objects that accompany them—also change with age. Age effects may actually reflect changes in motor development. Second, interactions emerged in our sample when modeling the effect of motor development on infants' looks, but not when modeling the effect of age. Finally, studies using an age-held-constant design, comparing infants of the same age who differed on motor abilities, have observed that motor development is associated with changes in object perception

(e.g., Soska et al., 2010; Rakison and Krogh, 2012; Libertus et al., 2013; Ross-Sheehy et al., 2016). Thus, although we cannot completely rule out age effects other than those attributable to motor development, it seems likely that our findings reflect, at least in part, change in motor ability.

Because our data were collected using families' own computers and webcams, it was necessarily more variable in quality than data collected in the lab. However, we demonstrate here that the quality of the data collected allowed for nuanced and in-depth data analysis at the level of infants' individual looks. Although looking coded from video does not have the temporal resolution of eye tracking data, we were nevertheless able to examine infants' behavior at multiple levels, gaining deeper insight into their looking behavior and how their visual preferences ebb and flow over time.

Specifically, our results indicate that infants' preference for the handled object—and differences between infants of different motor levels—occurred at the level of their individual looks. Early looks by more motorically advanced infants, (i.e., those who could crawl, stand, or walk) were longer to handled than to non-handled objects; this difference decreased over time. Thus, the Lookit platform, or other tools for online infant data collection, can generate the quality of data that allows researchers to ask sophisticated questions about the nature of infants' looking behavior and how it changes not only over development, but also from moment to moment during a trial.

Our analytic approach also allowed us to control for various potential confounding variables. Each of our effects of interest were obtained in analyses controlling for differences in object salience and object size. Thus, although infants generally looked longer to more salient objects, the effects of handled vs. non-handled were obtained in analyses that controlled for these potentially confounding factors.

This study was conducted online out of necessity due to the COVID-19 pandemic. However, the results contribute to our growing understanding of how motor development is related to infants' object perception, adding novel findings to the work showing such relations. In addition, we demonstrate how data obtained via online platforms can be effective in conducting sophisticated analyses that provide insight beyond overall preferences for one stimulus over another. Thus, online testing is an important avenue for future research in infant development.

DATA AVAILABILITY STATEMENT

The data and R scripts are available in this repository: <https://osf.io/a4ms6/>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by UC Davis Institutional Review Board

Administration. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study. Instead, parents provided a video recording of their consent.

AUTHOR'S NOTE

Research materials, statistical analyses, **Supplementary Materials**, and data sets are available from OSF (https://osf.io/a4ms6/?view_only=8974f5297e25457fa38ee815d7f760fe) and videos of subject sessions are available at Databrary (<https://nyu.databrary.org/volume/1316>).

AUTHOR CONTRIBUTIONS

CN and LO contributed to conception and design of the study. CN collected the data, conducted the data analysis, and wrote the first draft of the manuscript. LO contributed to the data analysis and wrote sections of the manuscript. Both authors contributed to manuscript revision, read, and approved the submitted version.

REFERENCES

- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw. Articles* 67, 1–48. doi: 10.18637/jss.v067.i01
- Baumgartner, H. A., and Oakes, L. M. (2013). Investigating the relation between infants' manual activity with objects and their perception of dynamic events. *Infancy* 18, 983–1006. doi: 10.1111/inf.12009
- Cicchino, J. B., and Rakison, D. H. (2008). Producing and processing self-propelled motion in infancy. *Dev. Psychol.* 44, 1232–1241. doi: 10.1037/a0012619
- Corbetta, D., DiMercurio, A., Wiener, R. F., Connell, J. P., and Clark, M. (2018). "Chapter one - how perception and action fosters exploration and selection in infant skill acquisition," in *Advances in Child Development and Behavior*, Vol. 55, ed J. M. Plumert (London: Academic Press), 1–29. doi: 10.1016/bs.acdb.2018.04.001
- Franchak, J. M. (2018). Changing opportunities for learning in everyday life: infant body position over the first year. *Infancy* 24, 187–209. doi: 10.1111/inf.12272
- Frick, A., and Möhring, W. (2013). Mental object rotation and motor development in 8- and 10-month-old infants. *J. Exp. Child Psychol.* 115, 708–720. doi: 10.1016/j.jecp.2013.04.001
- Gerhard, T. M., and Schwarzer, G. (2018). Impact of rotation angle on crawling and non-crawling 9-month-old infants' mental rotation ability. *J. Exp. Child Psychol.* 170, 45–56. doi: 10.1016/j.jecp.2018.01.001
- Harel, J., Koch, C., and Perona, P. (2007). "Graph-based visual saliency," in *Advances in Neural Information Processing Systems 19 (NIPS 2006)* (Cambridge, MA: MIT Press), 545–552.
- Horst, J. S., and Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) database: a collection of novel images for use in experimental research. *Behav. Res. Methods* 48, 1393–1409. doi: 10.3758/s13428-015-0647-3
- Iverson, J. M. (2021). Developmental variability and developmental cascades: lessons from motor and language development in infancy. *Curr. Dir. Psychol. Sci.* 30, 228–235. doi: 10.1177/0963721421993822
- Karasik, L. B., Tamis-LeMonda, C. S., and Adolph, K. E. (2011). Transition from crawling to walking and infants' actions with objects and people. *Child Dev.* 82, 1199–1209. doi: 10.1111/j.1467-8624.2011.01595.x
- Kretch, K. S., Franchak, J. M., and Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Dev.* 85, 1503–1518. doi: 10.1111/cdev.12206

FUNDING

This research and preparation of this manuscript was made possible by NIH grant R01EY030127 awarded to LO.

ACKNOWLEDGMENTS

We thank the students and staff at the Infant Cognition Lab at UC Davis, in particular Rebecca Beaton, Gabriela Ganoza, and Franchesca Quintero, for their help with data collection and coding. We express our appreciation to Aaron Beckner and Michaela DeBolt for their advice and help with the analyses and providing us with R code that provided the foundation of our data processing, and to Aaron Beckner, Gabrielle Blanch, Michaela DeBolt, Shannon Klotz, and Van Pham for comments on drafts of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.733218/full#supplementary-material>

- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13
- Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (2018). Emmeans: estimated marginal means, aka least-squares means. *R Package Version* 1:3.
- Libertus, K., Gibson, J., Hidayatallah, N. Z., Hirtle, J., Adcock, R. A., and Needham, A. (2013). Size matters: how age and reaching experiences shape infants' preferences for different sized objects. *Infant Behav. Dev.* 36, 189–198. doi: 10.1016/j.infbeh.2013.01.006
- Libertus, K., and Hauf, P. (2017). Editorial: motor skills and their foundational role for perceptual, social, and cognitive development. *Front. Psychol.* 8:301. doi: 10.3389/fpsyg.2017.00301
- Libertus, K., and Needham, A. (2011). Reaching experience increases face preference in 3-month-old infants. *Dev. Sci.* 14, 1355–1364. doi: 10.1111/j.1467-7687.2011.01084.x
- Libertus, K., and Needham, A. W. (2010). Teach to reach: the effects of active vs. passive reaching experiences on action and perception. *Vision Res.* 50, 2750–2757. doi: 10.1016/j.visres.2010.09.001
- Lüdtke, D., Ben-Shachar, M., Patil, I., Waggoner, P., and Makowski, D. (2021). Performance: an R package for assessment, comparison and testing of statistical models. *J. Open Source Softw.* 6:3139. doi: 10.21105/joss.03139
- Needham, A. W., Barrett, T. M., and Peterman, K. (2002). A pick me up for infants' exploratory skills: early simulated experiences reaching for objects using "sticky" mittens enhances young infants' object exploration skills. *Infant Behav. Dev.* 25, 279–295. doi: 10.1016/S0163-6383(02)00097-8
- Oakes, L. M., DeBolt, M. C., Beckner, A. G., Voss, A. T., and Cantrell, L. M. (2021). Infant eye gaze while viewing dynamic faces. *Brain Sci.* 11:231. doi: 10.3390/brainsci11020231
- Oakes, L. M., and Rakison, D. H. (2019). *Developmental Cascades: Building the Infant Mind*. Oxford: Oxford University Press. Available online at: <https://market.android.com/details?id=book-k3KfDwAAQBAJ> (accessed July 1, 2019).
- Perone, S., Madole, K. L., Ross-Sheehy, S., Carey, M., and Oakes, L. M. (2008). The relation between infants' activity with objects and attention to object appearance. *Dev. Psychol.* 44, 1242–1248. doi: 10.1037/0012-1649.44.5.1242
- Piper, M. C., Pinnell, L. E., Darrah, J., Maguire, T., and Byrne, P. J. (1992). Construction and validation of the Alberta Infant Motor Scale (AIMS). *Canad. J. Public Health* 83, S46–S50.

- Pomaranski, K. I., Hayes, T. R., Kwon, M. K., Henderson, J. M., and Oakes, L. M. (in press). Developmental changes in natural scene viewing in infancy. *Dev. Psychol.*
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: <https://www.R-project.org/>
- Rakison, D. H., and Krogh, L. (2012). Does causal action facilitate causal perception in infants younger than 6 months of age? *Dev. Sci.* 15, 43–53. doi: 10.1111/j.1467-7687.2011.01096.x
- Rochat, P., and Goubet, N. (1995). Development of sitting and reaching in 5- to 6-month-old infants. *Infant Behav. Dev.* 18, 53–68. doi: 10.1016/0163-6383(95)90007-1
- Ross-Sheehy, S., Perone, S., Vecera, S. P., and Oakes, L. M. (2016). The relationship between sitting and the use of symmetry as a cue to figure-ground assignment in 6.5-month-old infants. *Front. Psychol.* 7:759. doi: 10.3389/fpsyg.2016.00759
- Schwarzer, G., Freitag, C., and Schum, N. (2013). How crawling and manual object exploration are related to the mental rotation abilities of 9-month-old infants. *Front. Psychol.* 4:97. doi: 10.3389/fpsyg.2013.00097
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (part 2): assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/OPMI_a_00001
- Scott, K., and Schulz, L. (2017). Lookit (Part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Smith, L. B., Jayaraman, S., Clerkin, E., and Yu, C. (2018). The developing infant creates a curriculum for statistical learning. *Trends Cogn. Sci.* 22, 325–336. doi: 10.1016/j.tics.2018.02.004
- Sommerville, J. A., Woodward, A. L., and Needham, A. W. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cogn. Psychol.* 96, B1–B11. doi: 10.1016/j.cognition.2004.07.004
- Soska, K. C., and Adolph, K. E. (2014). Postural position constrains multimodal object exploration in infants. *Infancy* 19, 138–161. doi: 10.1111/inf.12039
- Soska, K. C., Adolph, K. E., and Johnson, S. P. (2010). Systems in development: motor skill acquisition facilitates three-dimensional object completion. *Dev. Psychol.* 46, 129–138. doi: 10.1037/a0014618
- Yamamoto, H., Sato, A., and Itakura, S. (2019). Transition from crawling to walking changes gaze communication space in everyday infant-parent interaction. *Front. Psychol.* 10:2987. doi: 10.3389/fpsyg.2019.02987

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Nelson and Oakes. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Organizing the Methodological Toolbox: Lessons Learned From Implementing Developmental Methods Online

Jonathan F. Kominsky^{1,2*}, Katarina Begus^{1,2}, Ilona Bass^{1,2,3}, Joseph Colantonio², Julia A. Leonard^{4,5}, Allyson P. Mackey⁴ and Elizabeth Bonawitz¹

¹ Graduate School of Education, Harvard University, Cambridge, MA, United States, ² Department of Psychology, Rutgers University, Newark, NJ, United States, ³ Department of Psychology, Harvard University, Cambridge, MA, United States, ⁴ Department of Psychology, University of Pennsylvania, Philadelphia, PA, United States, ⁵ Department of Psychology, Yale University, New Haven, CT, United States

OPEN ACCESS

Edited by:

Natasha Kirkham,
Birkbeck, University of London,
United Kingdom

Reviewed by:

Ola Ozernov-Palchik,
Massachusetts Institute
of Technology, United States
Przemyslaw Tomalski,
Institute of Psychology, Polish
Academy of Sciences, Poland

*Correspondence:

Jonathan F. Kominsky
jkominsky@g.harvard.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 29 April 2021

Accepted: 23 August 2021

Published: 13 September 2021

Citation:

Kominsky JF, Begus K, Bass I,
Colantonio J, Leonard JA, Mackey AP
and Bonawitz E (2021) Organizing
the Methodological Toolbox: Lessons
Learned From Implementing
Developmental Methods Online.
Front. Psychol. 12:702710.
doi: 10.3389/fpsyg.2021.702710

Adapting studies typically run in the lab, preschool, or museum to online data collection presents a variety of challenges. The solutions to those challenges depend heavily on the specific questions pursued, the methods used, and the constraints imposed by available technology. We present a partial sample of solutions, discussing approaches we have developed for adapting studies targeting a range of different developmental populations, from infants to school-aged children, and utilizing various online methods such as high-framerate video presentation, having participants interact with a display on their own computer, having the experimenter interact with both the participant and an actor, recording free-play with physical objects, recording infant looking times both offline and live, and more. We also raise issues and solutions regarding recruitment and representativeness in online samples. By identifying the concrete needs of a given approach, tools that meet each of those individual needs, and interfaces between those tools, we have been able to implement many (but not all) of our studies using online data collection during the COVID-19 pandemic. This systematic review aligning available tools and approaches with different methods can inform the design of future studies, in and outside of the lab.

Keywords: developmental psychology, online studies, metascience, behavioral methods, infant, early childhood

INTRODUCTION

In many ways, the COVID-19 pandemic has accelerated technological trends in psychological research, such as the use of online data platforms to carry out “research at scale.” Developmental research has tended to lag behind in adopting these alternatives, likely due to the demanding methodological sensitivities required for child participants. Nonetheless, the health-safety issues of the past year have forced developmentalists to confront these methodological challenges and

consider safer alternatives to in-person studies. This has revealed myriad potential advantages to online developmental research. Online research may enable labs to recruit more diverse samples, reduce barriers for participation compared to coming into the lab, facilitate longitudinal research by allowing for easier repeated access to the same participants, save researcher time by automating data collection, allow for naturalistic data collection, and more (Sheskin et al., 2020; see also Lourenco and Tasimi, 2020). Thus, there is ample reason to continue conducting developmental research online even after the COVID-19 pandemic has passed. The focus in this paper is to highlight the methodological lessons of this past year, to create a framework to help other researchers understand their methodological needs, and to identify available solutions for running developmental studies online.

Our methodological experiences are not necessarily novel. In the years leading up to the COVID-19 pandemic, a handful of developmental researchers were already pioneering various techniques for running experiments with children over the internet, without having to bring them into the lab (e.g., Scott and Schulz, 2017; Sheskin and Keil, 2018; Rhodes et al., 2020). However, once the pandemic hit, in addition to existing tools and techniques being used much more heavily, a number of new tools and techniques were quickly devised and put into practice. Because of the speed and urgency of this development, there are few compilations of the different techniques that different labs came up with, or the rationales behind why different techniques were used. To help researchers identify the best tools to conduct their developmental research online, we focus on a framework that starts with identifying the methodological constraints of a specific study, and we then present the available tools that meet those constraints. In addition, we consider the potential limitations or issues that these different approaches introduce and suggest ways to address those problems. We also discuss issues with recruitment and data quality that may arise with different approaches. In this way we hope to ‘organize the methodological toolbox,’ providing an easy reference for researchers to use when designing new studies in order to figure out how best to implement a given study online. The goal of this particular manuscript is to provide a how-to guide, rather than a comprehensive comparison between online and in-person methods (though we believe such comparisons should be a high priority for research in the coming years).

In the first part of this paper, the authors present six case studies from our own research methodologies, in order to give a general sense of the different kinds of approaches that are available, and the different kinds of studies that can be run. These case studies cover a wide range of approaches, from a very direct translation of an in-person task to online, to studies that allow the experimenter to take advantage of the unique properties of videoconferencing, to studies where there is no experimenter at all, and data collection is fully automated. In each case, we describe the goals and measures the study used, the methodological constraints and the approach used to meet those constraints, and any notable problems that needed to be addressed during the study. Furthermore, we have collected examples and guides of each of the approaches

used in these case studies in an OSF repository¹, to provide concrete examples for researchers interested in using these techniques in their own research. While each of these case studies comes from investigations of cognitive development, the techniques described may be generally applicable to many areas of developmental research. In the second part, we abstract away from these case studies in order to examine different methodological constraints that might arise in the design of a developmental study, and specific solutions that are available to address those constraints, with special attention to the pros and cons of different approaches. We also briefly consider issues related to the demographics of online populations and barriers to participation, although these issues have already received far more extensive consideration in other work (Lourenco and Tasimi, 2020; Sheskin et al., 2020).

CASE STUDIES

Case Study 1: Direct Translation of an In-Person Study to Online

This project (Kaminsky et al., 2021b) started before the pandemic and was adapted for online data collection. In person, the project involved showing participants (4-year-olds) a Qualtrics (2005) survey loaded onto a tablet. In the survey, participants first saw a short training about what an “x-ray” was, and then were shown two videos. In one video, a fur- or feather-covered puppet moved back and forth across a stage in an apparently self-propelled manner. In the second video, the other puppet (whichever was not in the first video) was shown sitting in a pink tray, being moved back and forth across the stage. It was important, particularly in the self-propelled case, that the movement appear smooth and not jerky. After each video, participants were asked to choose which of three images showed the “insides” of that puppet. In-person, they simply tapped the image on the tablet.

It was possible to directly translate this study to online data collection, with only two major methodological constraints. First, we needed a way to implement the multiple-choice response method. Second, we needed a way to present the videos such that the movement of the objects would look smooth. The multiple-choice response method was straightforward. An existing solution from the Yale Cognition and Development Lab is perfect for this kind of paradigm: simply present each of the options on a different color background, and train participants to respond by naming the color of their choice (Sheskin and Keil, 2018). We had used this technique in an earlier project that was also run online, prior to the pandemic (Kaminsky et al., 2021a). This response method avoids the problem of trying to decipher where children are pointing through a webcam, or figuring out a way to let them interactively click on a choice. This approach proved to be highly effective in this case: every one of the 30 participants run using this online method provided usable data.

For the video presentation, we found that Zoom screen-sharing was simply inadequate. At the start of the pandemic in particular, before the platforms underwent a substantial amount

¹<https://osf.io/g42rw/>

of development, screen-sharing had a framerate around 10 fps, and the graphical quality was such that the fur and feather textures of the objects became amorphous blobs of color. In order to present the stimuli smoothly and in high visual detail, we needed a solution that did not stream them from the experimenter's computer, but instead downloaded them directly onto the participant's computer, while allowing the experimenter to control when they were presented. The only system we found for this that could work on any computer operating system was a website called Slides.com, which allows you to create a slide-show, send the participant (or audience) a link, and then as you advance through the slides from the presenter account, the slides advance in the audience's web browser as well. Videos are presented as HTML5 video tags, which are downloaded in the participant's web browser and rendered on their own computer, meaning the video plays at its native resolution and framerate. slides.com is also free in a limited capacity, and a relatively inexpensive \$7/month if you need to store more presentations or use certain advanced features, but we have found that the free account provides all the functionality required. The only notable downside is that it was impossible to truly randomize the order of presentation of either the choices or the trials. Rather, we had to manually construct multiple pre-randomized orders in new slide decks and assign participants to them in advance of starting the online session.

Case Study 2: Processed Video Feed Over Zoom With Open Broadcaster Software

In another lab study investigating young children's social inferences, we wanted to know whether 6- to 8-year-olds would calibrate decisions selecting from recommended tasks based on an instructor's (false) beliefs about their competence (Bass et al., 2021). To this end, we designed an experiment in which a confederate (the "Teacher") overestimated, underestimated, or accurately represented participants' performance on a picture-matching game (between subjects). Using her "prior knowledge" of the participant's ability, this Teacher then presented three new matching games and evaluated them as much too difficult, not difficult enough, or just right for the participant; children then ranked their preferences for which of these new games they want to play. Children's verbal responses were coded into a spreadsheet by an experimenter in real-time, and this coding was checked with video recordings of the Zoom call by an independent coder after the study session.

Because this task would necessarily involve multiple testers (the experimenter and the "Teacher" confederate), coordinating schedules and technical setups would be difficult. Further, these studies must be carefully controlled across conditions, such that any experimental manipulations are delivered in exactly the same way every time, with no possibility for bias. To circumvent these potential issues, we instead opted to have only one live experimenter administer the task; we used pre-recorded videos of the Teacher to present to participants during the experiment, under the pretense that she was actually live in the call. The key piece of software used for this study was a program called

OBS, or "Open Broadcaster Software²." Using OBS, we were able to create a processed video feed that incorporated the experimenter's webcam and pre-recorded videos of the Teacher. This approach is less bandwidth-reliant than screen-sharing, allowing for higher framerate and resolution. By presenting this video feed over Zoom (along with some carefully timed acting from the experimenter "in response" to the pre-recorded Teacher), we created the illusion that the Teacher was also live on the call and interacting with the experimenter. (For a demonstration of how to execute the acting and timing as the experimenter, see: <https://osf.io/3r5cj/>. For a full video of a child being run in this task, and an example of what this set-up ultimately looks like to the participant, see: <https://osf.io/a4be7>. For a tutorial on how to use this set-up, see: <https://osf.io/8ycnf/>.) Importantly, the task itself was quite complex for children: It involved recursive mental state reasoning (i.e., "I know that you know that I know. . ."), contextualizing pedagogical actions given a second-order false belief, and calibrating subsequent decision-making to that false belief. Nevertheless, children appeared to be sensitive to our experimental manipulation, even using this online, pre-recorded paradigm. Only three (out of 60) children's data had to be dropped and replaced: two for failure to pass built-in memory checks, and one for terminating the task early. No data were dropped due to technical difficulties.

The prospect of being able to run such nuanced social cognitive reasoning tasks online is an exciting one, but there are also limitations to this approach. First, running these studies smoothly puts non-trivial hardware demands on the experimenter: We have found that the minimum specifications require a 2.3 GHz dual-core Intel i5 CPU and 16 GB of RAM. Second, there is a significant amount of preparatory work involved in recording and editing the pre-recorded videos, and in setting up the stimuli in OBS. (For the stimuli used in the OBS "scenes" for this study³ > Processed video feed over Zoom with OBS > Calibration to Teachers' Knowledge > OBS scenes - materials.) Third, all experimenters need to be quite comfortable with acting. For instance, the timing required to make the "conversation" between the Teacher and the experimenter convincing was quite precise; and without the use of additional plugins (e.g., Voicemeeter Potato⁴), the experimenter is actually unable to hear any audio from the videos played through OBS, making this timing even more difficult. We also had to explain away the Teacher's inability to interact with the participant; therefore, the experimenter and the Teacher had to feign surprise at an "unexpected technical glitch" that supposedly prevented the Teacher from being able to hear the participant. Indeed, this raises another perhaps obvious limitation of this approach: The pre-recorded actor cannot respond live to participants, which may decrease believability that the Teacher is actually live in the call. Finally, this task is quite long: Under ideal circumstances, it takes about 20 min, but it often runs longer than this. In addition to typical reasons that an online study might run long (e.g., parents requiring extra time to ensure Zoom is set

²<https://obsproject.com/>

³<https://osf.io/rzh9d/files>

⁴<https://www.vb-audio.com/Voicemeeter/potato.htm>

up correctly), in this study in particular, children would often engage the experimenter with thoughts about how to “fix” the Teacher’s audio glitch so that they could correct her false belief about their competence, lengthening the overall time it took to administer the task. Armed with knowledge of these limitations, however, we believe this online approach represents a promising way of assessing children’s social cognitive development, even when experimental manipulations are quite subtle and task complexity is high.

Case Study 3: Remote Investigation of Curious Play

Many studies in our lab require measuring children’s autonomous play with toys. One such approach, the “Novel Apothecary Box” task, was designed to measure 4- to 8-year-old children’s curiosity through their playful exploration of a box with many possible drawers (each containing unknown items). Specifically, the design of this task aims to quantify children’s exploratory behaviors similar to past studies of novel toy exploration (e.g., Bonawitz et al., 2011) via their discovery of a (bounded) set of unknown options, objects, or functions, in a naturalistic, play-like scenario.

The main challenge in designing the Novel Apothecary Box was determining how to emulate children’s experience of naturalistic play with toys. We were concerned that tablet-based interactions would not capture the life-like, proprioceptive experience of play, but also concerned about the feasibility and health risks of mailing a large toy to families and requesting return of materials following completion of the study. We thus devised a modified Apothecary task, in which participating families are mailed a “cheaper version” of our task, which they are then able to keep as thanks for participating in the study. Families received a package with sixteen envelopes containing different kinds of enclosed (inexpensive) small toys and play materials, split up into four color-coded categories of related items (e.g., magnetic items in the blue envelopes; pretend-play items in the yellow envelopes⁵). This package also contains written instructions for the child’s caregiver⁶, describing the purpose of the package and its contents for the family while also prompting them to wait for further instructions from the experimenters before opening envelopes or showing materials to the child participant.

The procedure itself is administered over video call (e.g., Zoom) with a live experimenter. After the experimenter ensures that the child’s webcam adequately captures the child’s hands and a playspace surface (approximately four-to-six square feet on the floor), the experimenter provides a prompt, highlighting one of the four color sets. Then, the child is free to explore the envelopes (and their contents) for up to 6 min, or until they notify the experimenter that they are finished playing. Play sessions are video recorded through the video call software.

From these video recordings, various aspects of children’s play are coded, including important dependent variables as measured in past studies on exploratory play such as: the amount of

time children play with the envelopes and their contents, the amount of time children spend playing specifically with the demonstrated (blue) category, the number of envelopes children open (total and per color), the order in which children open and interact with each envelope and their contents, and the number of unique combinations of objects children try during their play. Importantly, this task design allowed for finer control of potentially important perceptual aspects of the stimuli that may otherwise be lost (or face noisiness) due to differences in available technology and devices on the participant’s end (e.g., device brightness, volume, on-screen object sizes, potentially undisclosed device damage).

Limitations of the apothecary task come about due to its nature as a play experiment-at-scale. Pre-pandemic, exploratory play studies used to measure children’s playful curiosity may have only required the creation of two sets of stimuli (e.g., one novel toy for testing, a second identical novel toy as a backup). Given that each participating family requires identical stimuli to maintain control over the experiment, the number of stimuli sets scales linearly with the number of participating families. For example, across our various studies employing this method, we have mailed more than 200 identical packages that must be purchased (~\$5 cost of toys items), hand-packed by participating experimenters, and mailed to families (~\$5 shipping). To mitigate the required labor in preparing packages, the items and packing materials were chosen in their simplest forms (single items in single envelopes) and prepared in large batches (typically up to 20 packages per batch). Additionally, depending on the climates of the locations between the experimenters and participants, issues with postal services may arise. Currently, we have only experienced approximately 10 percent attrition (103 of 114 recruited participants provided usable data) in regard to families not receiving the Apothecary task stimuli in the mail. For those who failed to receive their package, another package would be promptly prepared and sent to the participating family. Furthermore, if the experimental session were scheduled with an expected delivery date that was not met, participating families would simply be rescheduled to a future time slot, if desired.

Case Study 4: Online Infant Habituation Studies Using PyHab

Experiments with infants and toddlers involve methods that mitigate developmental limitations on talking and acting. In a habituation study, both the order of trials and (typically) each individual trial are gaze-contingent (Colombo and Mitchell, 2009). A typical habituation study involves trials that end when the infant has looked at the stimulus for some amount of time, and then looked away for some amount of time or a maximum trial length has been reached. Habituation trials are presented repeatedly until a habituation criterion is met, typically something like a total gaze-on time during the most recent X trials that is some fraction of the gaze-on time in the first Y trials. This means that infants’ gaze behavior must be coded by the experimenter in real time, so the experimenter can end a trial at an appropriate time,

⁵<https://osf.io/sc4rt/>

⁶<https://osf.io/t89yp/>

present the next trial, and determine when to proceed from habituation to test trials.

We were conducting a habituation study with 6- to 7-month-old infants in which the stimuli required smooth framerates, and the procedure required live gaze coding in order to determine when infants were habituated and when each trial should end. In the lab, these studies were run with PyHab (Kominsky, 2019), an add-on for PsychoPy (Peirce et al., 2019). To adapt them for online use, we took advantage of PyHab's open-source nature and modified it such that we were able to integrate it with a solution used in Case Study 1 (above) for smooth remote stimulus presentation: slides.com. In short, this modified version of PyHab controls a Slides.com presentation instead of directly presenting videos as it does in the lab. The parent of the participant is asked to open the slides.com presentation in a web browser and make it full-screen, so it is the only thing the infant can see, and then sit in such a way that the infant is visible on the webcam in Zoom. The experimenter then mutes themselves and watches the infant through the Zoom call, live coding whether the infant is looking at the screen or not, and PyHab determines when to end a trial and when to advance from habituation to test.

In many regards, once configured, the methodological experience is almost identical to running a habituation study in the lab, particularly if the experimenter is already familiar with using PyHab for in-lab studies. The initial setup is very different, however, and does require a small degree of technical skill to modify PsychoPy to interface with a web browser. We created a detailed step-by-step setup guide to help researchers do this setup more easily. This guide can be found at <https://osf.io/g42rw/>. In data collection to date we have had to exclude 2 of 17 participants, both due to environmental distractions (pets or siblings). Additional concerns regarding camera placement, home-based testing environments, and parental interference are discussed below.

Case Study 5: Unmoderated Online Study of Toddlers' Predictive Looks

Additional studies in our labs involve measuring infant looking behavior using eye-tracking and measuring concurrent brain activity, using EEG (electroencephalogram, measuring electrical activity recorded on the scalp, using specialized "nets" and software.) One such study was started before the pandemic, and originally involved EEG and eye tracking measures. Although it was impossible to move to an "online EEG" set-up, one aspect of the study could be salvaged. That is, one of the dependent variables of interest was whether toddlers would produce predictive saccades toward certain locations on the screen which would indicate that they have learned a rule. We define a 'predictive look' as an eye-movement toward specific locations on the screen, during a specific period of the trial, which is not elicited by any changes in the visual stimuli itself (the scene is static), but can be presumed to be driven by the participants' expectation of how the events will unfold. Specifically, if participants learn that certain objects get placed in one location, and another type of objects in another location, we can test whether participants anticipate where an

object will be placed, by examining whether they would saccade toward the correct locations, even when the placement does not in fact happen.

The study was adapted for online data collection using the platform Lookit (Scott and Schulz, 2017), developed by MIT Early Childhood Cognition Lab. Lookit offers experimenters a detailed tutorial and support on how to set up an online study, and offers participating families the possibility to take part in studies from home, at a time of their choosing, requiring only a computer device with a webcam. Participating in a study involves the caregiver reading or watching customized video instructions, created by the experimenter for the specific study, explaining the aim of the study, the duration, and the ideal set-up for optimal data collection. Video consent is obtained for each participant, and reviewed by the experimenter before access to the participant's video recording is obtained.

Adapting an in-lab toddler experiment to an unmoderated online experiment introduces some challenges. Participants' homes inevitably mean a less controlled environment for data collection than in-lab studies. In order to minimize the likelihood of losing data due to disruptions, poor video quality, or parental interference, detailed instructions with visual displays of how to participate are essential. Examples of video instructions for the participating families used in this study can be found here (<https://osf.io/6f5dj/> and <https://osf.io/9ep2t/>).

Another issue is calibrating gaze position. The outcome measure of interest in this study is toddlers' (15–18 months) predictive looks toward specific locations on screen. As opposed to in-lab studies – in which all participants would see the stimuli on the same screen, at the same distance, and positioned centrally with respect to the screen – self-administered online studies introduce variability in these parameters. To account for this variability and to maximize the reliability of analyzing toddlers' looking behavior, we introduced 'calibration' videos immediately preceding each of the test videos. In these videos, a captivating animation is displayed against a black background, at each of the crucial parts of the screen sequentially (corresponding to the locations toward which predictive looks are expected). This was followed by a centrally displayed animation, to bring the toddlers' attention back to the center of screen. These calibration videos allow the experimenter to establish what the participants' eyes look like when they fixate each of crucial locations of the screen, and therefore facilitate accurate coding of predictive looks in the following test videos, even if the participant is not sat centrally or their head position is not upright and forward facing. An example of a calibration video (followed by test video) used in this study can be found here (<https://osf.io/uvdjf/>).

Finally, while in-lab equipment typically allows for combining video recording of the participant with the display of the stimuli that the participant is watching, Lookit recordings only include the video of the participant. This means that in order to track the progression of the stimuli that the participant is observing, the experimenter must rely on the audio of the recording. This poses a particular challenge, if the experiment involves many trials and the audio of the videos is identical across trials (as was the case in our study). Experimenters should be conscious of this constraint when designing an experiment and add audio cues

(such as the calibration videos used before test trials in this study) to facilitate easier decoding of the recordings. Note that despite these precautions, some of the video recordings obtained through Lookit for this study did not contain the audio of the presented stimuli. It appears that certain webcams only record the audio coming from the environment, while filtering out the sound that is emitted by the device itself. This is an issue that, to the best of our knowledge, does not yet have a solution. It may therefore be good practice to design stimuli in ways that the illumination of the screen changes significantly (i.e., at the beginning of test trials), so that the reflection of this change may be detectable on the recordings of participants' faces and used for coding.

Case Study 6: Unmoderated Tablet-Based Game

This project was started before the pandemic. In its original form, children played with a physical wooden tree and used a pulley device to get an "egg" (a metal ball) back to a nest in the tree. Unbeknownst to the child, there was an electromagnet in the pulley device that allowed the experimenters to surreptitiously control when the egg fell off. We assigned children to conditions where the egg fell off at continuously closer positions to the nest or at about the same position each time. We were specifically interested in how children's trajectory of past performance influenced their decision to keep playing with the current tree or switch to an easier, shorter tree.

We had the following criteria for a remote version of this study: (1) asynchronous data collection to avoid scheduling and internet issues, (2) interactive design where children could feel like they had agency over their play, and (3) parent supervision that could ensure data quality with 4–6-year-olds, but was not intrusive. Based on these criteria, we concluded that the best solution would be to build an interactive touch screen web-based game for children. A undergraduate research assistant with strong coding skills built the game with the JavaScript library React⁷, hosted on Heroku⁸, and used MondoDB⁹ for the database.

The web-based version of the game was fairly similar to the in-person version. Children still had to get an egg back to a tree, but this time used their finger to slide up the platform with the egg instead of using a physical pulley device. The egg wobbled as it went up and fell off at predetermined points. As in the in-person game, at the end, children chose whether to keep playing with the current tree or switch to an easier, smaller tree. We wanted parents to supervise their child's play in case anything went wrong (child clicks wrong thing or closes game), but did not want parents to intervene. To this end, we instructed parents to quietly watch their child play and only answer questions in the game addressed to them ("parent, please confirm that child pressed X"). To ensure successful remote administration of the game, we had explicit instructions for the parent and child throughout. For example, we used pictures and verbal prompts to instruct children when to put their hands on their lap and listen and when to touch the screen and play. We also figured out

that design features could serve as implicit instructions: we only displayed the egg on the screen when children were supposed to move it around. To make sure a child was playing the game, and not an adult, we audio recorded participants' responses to questions about their name, age, and their final task choice using the npm module mic-recorder-to-mp3¹⁰. At the very end of the game, parents could also write in if there was any interference during game play.

We ran into three main issues with remote data collection: non-serious participants, game play issues, and voice recording problems. We had to halt our first round of data collection due to a large number of non-serious participants (~40%). The non-serious participants were unintentionally recruited through our Facebook ads and we only started to see them after we increased our compensation from \$5 to \$10 in hopes of attracting more participants. We spotted them thanks to the audio recordings. It turns out it is easy to tell an adult voice from a child voice – two people listened to recordings and always agreed when someone was an adult. We stopped the non-serious participants from participating in our study by halting payment and not inviting participants to play who had questionable information in their sign-up forms (e.g., different dates of birth entered on separate pages of questionnaire). Another issue we experienced was children unintentionally closing or restarting the game part-way through play. Because the game requires moving a pointer finger up the screen, it was easy for children's fingers to slip and refresh or close the page. Through the backend of the game, we received information on how many times children played the game and when they stopped playing. We usually followed up with parents directly via email to confirm details of their children's game play if it was halted early. We ended up excluding children who did not play through the full game in one session on the first try (7% of recruited participants excluded for this reason). However, the largest contingent of participants we had to exclude were those who did not have audio recordings (11% of recruited participants). We are unsure of the reason behind this issue and are continuing to investigate solutions.

Case Studies Summary

These six case studies illustrate a number of different approaches to conducting developmental research online, but this is far from a comprehensive list. Furthermore, the options that are available will certainly change as new technologies and services are developed. In the remainder of the paper, we will consider each of the columns in **Tables 1A, B**, including the factors that might go into each decision researchers can make in designing their experiment, and the tools that are available to experimenters based on those decisions. First, we examine issues of study design: how you construct your stimuli in the first place. Then we consider issues around actually running the study, i.e., the process of data collection. Finally, we discuss issues relating to the processing and analysis of data, including attrition, data reliability, and comparisons with in-lab data.

⁷<https://reactjs.org/>

⁸<https://www.heroku.com/>

⁹<https://www.mongoddb.com/2>

¹⁰https://github.com/Hunterzhaoliu/learning_curve

TABLE 1A | Features, advantages, and disadvantages of each of the six case studies.

| | Moderated | Stimulus fidelity | Setup effort | Technical requirements for implementation | Technical requirements to participate |
|--|-----------|--|--|--|--|
| Case 1: Direct translation of in-lab study | Yes | High – framerate and visual quality are rendered by participant computer and comparable to in-lab | Moderate – Building multiple presentation orders in slides.com | Low – No specific technical skills required, just experience building powerpoint-like systems | Low – Parent only needs web browser |
| Case 2: Processed Video Feed over Zoom with OBS | Yes | High – framerate and visual quality are rendered by participant computer and comparable to in-lab | Moderate to high – Installing and configuring OBS and all necessary plugins, recording and editing videos, setting up scenes in OBS | Moderate to high – Smooth video presentation requires 2.3 GHz dual-core Intel i5 CPU and 16 GB of RAM | Low – Parent only needs Zoom, and can use any device with a sufficiently large screen to see stimuli (tablet, desktop, laptop, etc.) |
| Case 3: Remote investigation of curious play | Yes | High – Stimuli are physical objects in the real world | Low to moderate – Minimal technical requirements as noted alongside proper participant cooperation for camera setup | Low to moderate – Current stimuli sets requires a minimal understanding of electronics | Low to Moderate – For behavioral coding purposes, a stable internet connection is required |
| Case 4: Online infant habituation studies using PyHab | Yes | High – framerate and visual quality are rendered by participant computer and comparable to in-lab | High – Configuring PsychoPy, setting up the slide show, inputting stimulus information into PyHab experiment | Moderate to high – Setup requires modifying PsychoPy with additional libraries, running PsychoPy, Zoom, and a web browser simultaneously | Low – Parent only needs web browser |
| Case 5: Unmoderated online study of toddlers' predictive looks | No | High – framerate and visual quality are rendered by participant computer and comparable to in-lab | High – arranging between-institution ethics agreement, coding the experiment on Lookit, recording and editing stimuli and instructions videos, peer review process of study. | Moderate – Setup requires video editing software for stimuli and instructions videos, and use of several online platforms – stimuli repository, experiment coding on Lookit, Slack for set-up support. | Low to moderate – Participation requires a device with a webcam and web browser, setting-up an account on Lookit, and following the set-up instructions and recording a video consent. |
| Case 6: Unmoderated tablet-based game | No | Medium – framerate and visual quality are rendered by participant computer, BUT dependent on participant's internet connection | High – coded game in JavaScript, handled database with MondoDB | High – Setup requires coding in JavaScript and general coding knowledge. | Low – Parent only needs web browser and touchscreen device (tablet, phone) |

The first column is descriptive. The remaining columns offer the experimenter's subjective opinion of different features of the methods.

DESIGNING A PROCEDURE: WHAT DO YOU NEED TO DO, AND HOW CAN YOU DO IT?

Here we will discuss different decisions researchers need to make in designing their studies. For easy reference for mapping decisions to tools, we provide a summary flow chart (**Figure 1**) which lists various solutions for different kinds of study design. However, in this section we also consider *why* you might choose to conduct a study in one way or another, to help researchers make informed decisions.

Moderated vs. Unmoderated: Do You Need an Experimenter?

We have found one of the most foundational decisions about online study design is whether the experimenter needs to be present during the experiment (i.e., “moderated”) or if the experiment can be run completely automatically, generally through a website of some kind (i.e., “unmoderated”). For some types of studies, an experimenter is absolutely necessary, including studies that examines how children interact with an adult as a primary question of interest, or infant habituation studies that requires gaze-contingent stimulus presentation (at least until automated gaze-coding technology becomes

substantially more advanced; Chouinard et al., 2018). For other types of studies, it is a choice, and there are advantages and disadvantages to each type of study.

For moderated studies like Case Studies 1–4, with a live experimenter, there are a number of clear upsides. First and foremost, a live experimenter can adapt better than an automated system to situations that might arise. An experimenter can ensure that data are being recorded correctly (e.g., the participant is visible on the webcam and can be heard), and the experimenter can be responsive to the participant's behavior in order to keep them engaged with the task. This is also relevant for tasks in which there are follow-up questions that are contingent on what the participant says. For example, while most automated systems can have a branching task structure, at least with our current technology, it would be difficult to have an automated task that reliably responded to a *verbal* response made by a participant. Second, moderated designs are more comparable to most in-lab studies. While there is value in replicating a study that was previously run by a live experimenter using an automated system (e.g., Scott and Schulz, 2017), if you are attempting to build on a previous finding and want to stick as closely to its methods as possible, running it with a live experimenter may be preferable.

The downsides of having a live experimenter are primarily that it requires scheduling an appointment and takes the

TABLE 1B | Features, advantages, and disadvantages of each of the six case studies with regard to data collection and analysis.

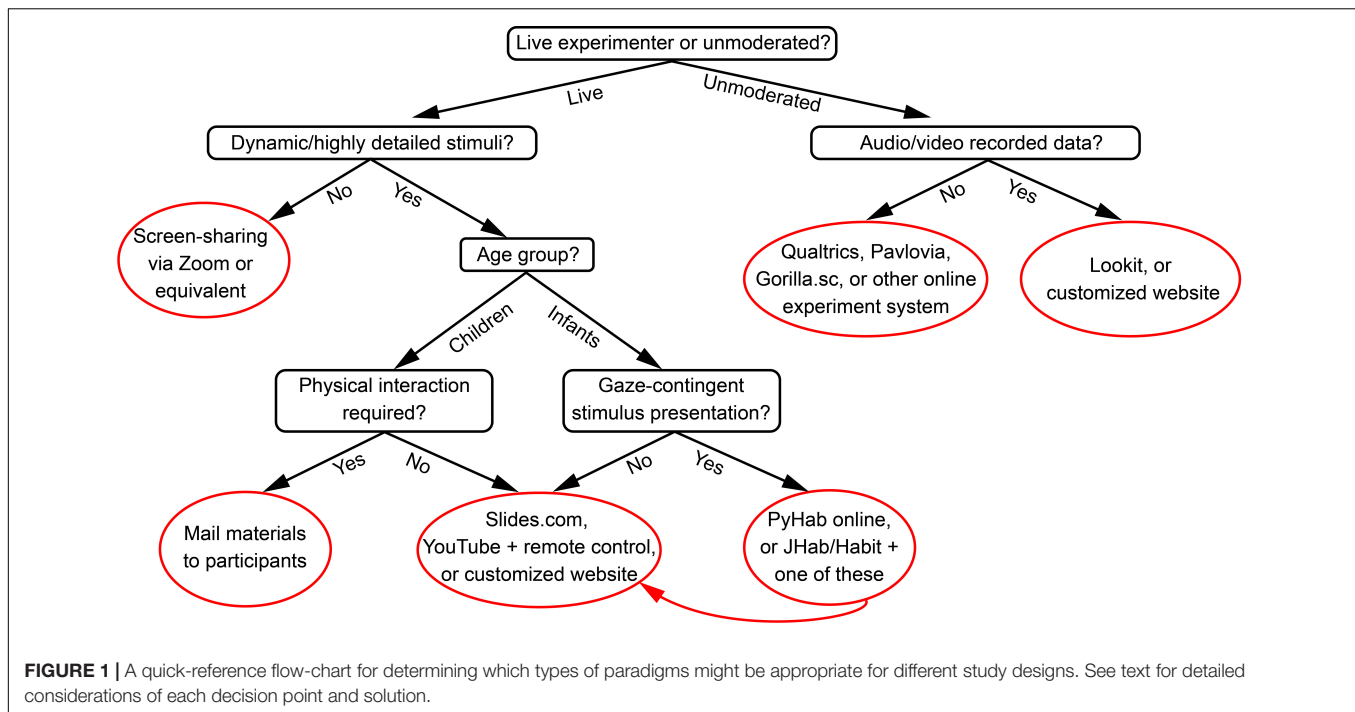
| | Running effort | Data processing effort | Attrition | Data reliability | Monetary cost |
|--|--|---|---|--|--|
| Case 1: Direct translation of in-lab study | Moderate – Similar to running a study in the lab, but with a browser and a Zoom window | Moderate – Data must be recorded outside of the presentation system, either manually or by coding video | Low – no online participants had to be excluded in this study | High – no difference between an in-lab and online sample. | Moderate – participant compensation in gift cards, potentially slides.com subscription |
| Case 2: Processed Video Feed over Zoom with OBS | Moderate – Similar to running a study in the lab, but with OBS, a browser and a Zoom window | Low to moderate – Data are manually coded by experimenter in real time, and checked with video recordings by an independent coder after the study session | Low – No higher than in-lab studies | High – Behavioral data from both adults and children were in line with our predictions | Low to Moderate – participant compensation in gift cards (all software is free) |
| Case 3: Remote investigation of curious play | Moderate to High – Data collection entails creation of multiple sets of physical stimuli, scaling with the projected sample size | Moderate – Data must be recorded/coded by a condition-blind researcher after the post-data collection session | Low – Participants and their families typically prepare a testing space adequately to ensure data means quality standards needed | Moderate to High – Behavioral data following similar trend to in-person samples. However, data collection and analysis is ongoing. | Moderate to High – Costs include compensation to participants, materials and labor for preparing packages, and shipping fees for delivering packages to participants |
| Case 4: Online infant habituation studies using PyHab | Moderate – Similar to running a habituation study in the lab, but with Zoom and PyHab | Low – Data recorded automatically by PyHab | Low – no higher than in-lab, and we are getting fewer fuss-outs from 6–7-month-olds | Moderate to High – Trends and SDs are thus far similar to in-lab, but data collection is ongoing. | Moderate – participant compensation in gift cards, potentially slides.com subscription |
| Case 5: Unmoderated online study of toddlers' predictive looks | None – data collected without involvement of experimenter. | Moderate – data is manually coded from video recordings by two independent and condition-blind researchers, post data-collection. | Moderate (~30%) – some video recordings could not be reliably coded due to audio issues, and poor positioning/visibility of the participants' eyes. | Moderate to High – prevalence of predictive looks similar to that found in-lab, using an eyetracker. | Low to Moderate – all online platforms are free to use. Experimenters can offer participating families compensation in the form of e-vouchers. |
| Case 6: Unmoderated tablet-based game | Low – data collection required emailing interested participants and paying them after participation | Moderate – voice recording data had to be manually checked and double entered. Confusing cases were discussed over email with parent. | Low to Moderate – excluded ~20% of data collected for issues with game play, audio recordings, or incorrect age. | High – no difference between an in-lab and online sample. | Moderate – compensation in gift cards, potentially paying someone to program the task. |

Each column offers the experimenter's subjective opinion of different features of the methods.

experimenter's time. Unmoderated studies are completely on the participants' schedule, while moderated ones require coordination between the participating family and the experimenter(s). Of course, that's also true of in-lab studies, and in fact moderated online studies are much easier to schedule and run than in-lab studies because they don't require anyone to travel. Furthermore, there are tools that can make signing up easier, such as using automated scheduling tools like Calendly or YouCanBookMe to allow parents to select a time that works for them, and these services often provide automated reminder emails that reduce no-shows. Anecdotally, we have found that an automated reminder email sent from the scheduling service 24 and 1 h before the appointment with a link to cancel or reschedule leads to very few unexpected no-shows (though we have not quantified the no-show rate precisely because people who do not show up for the study at all are not counted as participants). The other potential downside, which is again shared with typical in-lab studies, is that moderated studies introduce the possibility for experimenter effects or inconsistency between participants that unmoderated studies do not.

One of the advantages of unmoderated studies like Case Studies 5–6 is that, as mentioned, they do not require coordinating schedules between participant and experimenter. The participant can take part in the study at any time. Aside from just being easier, this also matters for studies that are trying to recruit from a global population: you don't need to worry about time zones. It also places zero burden on the experimenter, other than advertising the study and dealing with the data. For researchers who work with adults, the difference between running a study in the lab and running it over MTurk and Prolific is hard to overstate. A study that would take weeks or months in the lab often takes no more than a day through large online collection sites, all while the experimenter can be working on other things. In our experience, you don't typically get the same kind of pace of data collection with unmoderated developmental studies as you do with adult unmoderated online studies, but it is easier on the experimenter's schedule.

There are several drawbacks to unmoderated developmental studies, however. First, much careful thought needs to be put into how they are set up. Depending on the nature of the study, it may also require substantial technical skills to set up, involving



programming in JavaScript or even full web development. Furthermore, there is a real challenge in the “user experience” aspect of the study. The experimenter is not present to give instructions or correct anything the participant does, so the study must be thoughtfully designed to ensure that the participant completes the task as intended. Lookit (Scott and Schulz, 2017) has taken great pains to design rich instruction templates for this very reason (these are described in the Lookit tutorial¹¹), but if you are not using an established system, you would need to design your own. The lack of a live experimenter also means you may face challenges with data quality and attrition (see the “Data: Attrition and Quality” section below).

Based on these considerations, researchers should think carefully about whether a study is better served being moderated or unmoderated. That decision then constrains which tools are appropriate for the study.

When considering tools for moderated studies, the solution is almost always going to involve some kind of video-conferencing software such as Zoom, Skype, FaceTime, Google Meet, Adobe Connect, or others. There are different stimulus presentation systems that can be used alongside the video-conferencing software, depending on the needs of the stimuli (see next section), but the video-conferencing software is always how the experimenter interacts with the participating family. In principle one could also run a study over a phone call or with mailed surveys, but it would be much more restricted in terms of the types of data that could be collected, and the types of stimuli that could be presented.

For unmoderated studies, there are a variety of potential solutions, but the first major consideration is what kind of data

the researcher intends to collect. If video or audio recordings of the participant doing the study are needed, the available solutions are Lookit, or something custom-built that can access the participant’s webcam and/or microphone via the web browser (e.g., see Case 6 and the accompanying materials in the OSF repository at <https://osf.io/rzh9d/>). On the other hand, if it is sufficient to collect responses via keyboard, mouse, or touchscreen, there are several options. For simple survey-like studies that involve multiple choice elements, even with audio or video, there are services like Qualtrics, SurveyMonkey, and others. These systems often offer institutional licenses, so check with your university IT office about whether an online survey system is available to you. For studies that involve more complex stimuli or tasks, there are online psychophysical study presentation systems like PsychoPy’s Pavlovia, Gorilla.sc, Labvanced, jsPsych, Testable, OpenSesame, and others (for a careful examination of the stimulus presentation capabilities of many such systems, see Bridges et al., 2020). Of course, it is also possible to create a custom web app instead if the researcher has or has access to someone with the required technical skills. An additional option, which requires yet further technical skills and substantial effort, is to create a data collection platform for mobile devices (e.g., Kid Talk Scrapbook¹²). The use of mobile apps for developmental research is still new and relatively untested at time of writing, though there has been at least one pediatric medical study using a mobile app-based platform (Lalloo et al., 2020).

Finally, there is a sort of compromise category that is neither strictly moderated nor unmoderated: asking parents to serve as experimenters. Some unmoderated studies are effectively already like this, they ask the parent to monitor their child completing

¹¹<https://lookit.readthedocs.io/en/develop/tutorial-access.html>

¹²<https://www.kidtalkscrapbook.org/>

the task to keep them focused. In general, a study like this would share many of the advantages and disadvantages of an unmoderated study but might allow for some designs that would otherwise be impossible. For example, consider a study that focused on a particular daily routine and asked parents to record that routine, and ask specific questions during it, and then send those videos to the experimenter (Leonard et al., 2020). It is something of an edge case, but for certain research questions it may be the best approach.

Stimulus Presentation: Speed and Detail

Stimulus presentation introduces another important consideration for developmental studies. Particularly for moderated studies, a key concern is how high-fidelity the stimuli need to be. First, let us carve out an exception: studies like Case 3, in which physical stimuli are delivered to the participating family, are obviously the highest possible level of fidelity. If the research question involves children physically interacting with an object, this is obviously necessary, but the cost and logistical difficulties introduced by shipping materials to each individual participant are high. This section will mostly be concerned with screen-based stimuli.

In terms of screen-based stimulus quality, there is one key technical consideration: is the experimenter's computer rendering the stimuli and then streaming it to the participant's computer via screen-sharing of some kind, or are the stimuli being rendered on the participants' computers directly? Screen-sharing imposes some caps on the quality of stimuli in various ways. The resolution (number of pixels/level of detail) may be restricted, and for dynamic stimuli, the frame-rate may be reduced or unstable. Case 1, above, presents an example of stimuli that could not be used with screen-sharing. However, if the stimuli for an experiment are static images or otherwise do not lose relevant information if the video quality or frame-rate should happen to drop, there is no reason *not* to use screen-sharing. There are two advantages to screen-sharing over other solutions. First, it is the easiest way for the experimenter to control the stimulus presentation, because the stimuli are being displayed on the experimenter's own computer and that view is being sent to the participant. Second, it can be easier for the participant (or their parent) to set up, because it does not require them to open a separate web browser or other program in order to view the stimuli, just the video-conference they would have to open anyways.

There are also multiple ways to stream stimuli from an experimenter's computer over a video-conference, and different approaches can offer some methodological flexibility. For anyone who has used Zoom, the most obvious and simple solution is the built-in screen-sharing feature, and many other video-conferencing systems offer similar capabilities. In these cases, the image on the experimenter's screen is captured by Zoom and transmitted to the participant alongside the image captured from the experimenter's webcam. The frame-rate of screen-sharing like this is typically low, often capping out at 10–20 frames per second, subject to the upload speed of the experimenter's internet connection *and* the download speed of the participant's internet connection. An alternative solution is the one described

in Case 2, in which the experimenter uses an additional program to create a processed video feed that is treated as a “virtual webcam.” This can sometimes offer slightly higher-quality video performance because only one video feed is being transmitted instead of two, so it is less restricted by upload speed. However, the main advantage is that it allows for designs like the one described in Case 2, in which there is a pre-recorded additional experimenter ‘present’ on the video call in a way that looks convincing, and does not require an additional experimenter to actually join the video call.

However, in cases where the stimuli need to be higher-quality, the best solutions are going to be those that download the stimuli on to the participant's computer directly in some way and have the participant's computer render the stimuli at their native resolution and framerate. Unmoderated studies necessarily do this: the participant accesses a website which downloads the stimuli into their browser and renders the stimulus file at its native resolution and frame-rate. There are various ways to achieve this in a moderated study as well, but it typically involves asking the participant to open a web browser and navigate to a particular website where the stimuli are hosted. In Cases 1 and 4 above we describe one such system, Slides.com, which has the dual advantages of allowing the experimenter to control when the stimuli are presented and of being platform-universal (i.e., not restricted to Windows or Mac systems). However, it has other limitations, notably an inability to keep the experimenter blinded to the experimental condition or randomize presentation order on its own (though Case 4 works around this by having PyHab control the order of slides).

Another solution some researchers have used is asking participants to share *their* screen with the experimenter and using Zoom's ‘remote control’ function to allow the *experimenter* to control the stimuli *on the participant's computer* (Liu, 2020). Essentially, the participant hands over partial control of their computer to the experimenter. Alternatively, the experimenter can send the participant to the sort of website that would host an unmoderated study, like a Qualtrics survey or even a Pavlovian (or equivalent) experiment, and have them complete the study while talking to the experimenter. This provides a way to conduct an interactive task (i.e., in which the participant directly interacts with objects on the screen) with the advantages of a moderated study. Across all of these solutions, it is often worth asking the participant to share *their* screen with the *experimenter*, so that the experimenter can record what the participant is seeing. In Zoom, this also keeps the experimenter visible as a small window in the corner but allows the stimuli to take up the bulk of the screen.

One concern that can arise using systems that present stimuli through a web browser is whether the stimulus files are in a format that will work on the participant's computer. When screen-sharing, as long as the stimuli render on the experimenter's computer, they are fine, because what appears on the experimenter's screen is what the participant will see. For other presentation systems, the safest thing to do is use universal file formats. The safest file format to use is MPEG-4 (.mp4) made with h.264 compression, because these types of video files are supported by all major web browsers and operating systems as of 2021. For audio files, .wav files are safely universal, as are .mp3

files, though *creating* .mp3 files can be more difficult because it is a proprietary codec. In terms of making the stimulus files themselves, whatever solutions researchers have used in the past should still work, provided they can export to these standard file formats (and most audio and video editing software can do exactly that).

To sum up, what kind of solution researchers should use will depend on the level of visual quality your study requires, the nature of the stimuli, the level of interactivity required, and what solutions the researchers are most comfortable with from a technical perspective. In **Figure 1**, we summarize these considerations in what we hope will prove an easy reference for researchers figuring out what kind of tools to use for their online studies.

RUNNING STUDIES AND DEALING WITH DATA

Developmental studies must be sensitive to the abilities and nature of their participants. It would not make sense to design a study for 6-month-olds that required a verbal response, for example, or a study for 3-year-olds that required attending to a tedious task for 30 min. This is obviously still true when it comes to online studies, but there is an additional constraint that researchers should consider: the technical demands on the participant and their parents to participate in the first place. In general, researchers should strive to make an online study as easy as possible for participants to take part in. In other words, as much as possible, participating in an experiment should not require participants or their parents to need to conduct extensive technical setup, rely on parents using a specific operating system or web browser, or reconfigure the space in their home in which the experiment will be run. There are some specific cases where some of these might be unavoidable, for example a study that involved examining toddler's mobility behavior at home would require there be a sufficiently large space for them to move around in, but in general we should strive to make the barriers to participation as low as possible, especially given that merely having a computer, reliable internet connection, and time can all be barriers to many participants (Lourenco and Tasimi, 2020).

Of the solutions discussed in the previous section, none require the installation of specific software on the participant's computer beyond a web browser and video-conferencing software (which in many cases can run through a web browser anyways). It is our opinion that Lookit (Scott and Schulz, 2017) demonstrates a reasonable upper limit of what we can ask of parents, particularly for unmoderated studies, and Lookit asks as little as it can while still collecting usable data. The designers of Lookit have very carefully created a process that balances the demands on parents with the needs of experimenters. Participating in a study on Lookit requires no additional software or technology, but involves a multi-step setup in which parents are carefully walked through making sure their webcam and microphone are operative, recording a consent statement and test video, and making

sure the participating child or infant is properly located on the screen. This step-by-step guide is the absolute minimum that can be asked of parents to ensure they will be able to complete a Lookit study successfully, and its instructions have been carefully refined over the years Lookit has been in operation. (see also the Lookit 'getting started' guide for more information about this process: <https://lookit.readthedocs.io/en/develop/researchers-start-here.html>).

There are some hardware constraints on the participant for these studies as well. The most obvious ones are a computer with a microphone and webcam. Some studies can be conducted on mobile devices like tablets or smartphones, but not all. For example, the techniques described in Case Studies 1 and 4 would not work on a tablet because most tablets cannot simultaneously run Zoom and a web browser, or they cut off the webcam when the web browser is the focal app. Studies that are run entirely in Zoom like Case Studies 2–3, or custom-programmed web apps like Case 6, could be run on a participant's tablet or smartphone, at least in principle, though researchers should consider whether their particular study requires screens of a minimum size for effective stimulus presentation. At this time it is not possible to participate in a Lookit study (like Case 5) from a tablet, though future development could change that. More generally, depending on the nature of the stimuli and the study design, researchers should consider if there are minimum screen sizes or resolutions that would present difficulties. For perception studies that require more precise viewing conditions, there are techniques for asking participants to calibrate their screens using an object of standard size (for an example, see Bechlvianidis and Lagnado, 2016, Appendix A2). Even if that level of precision is not needed, it may be worth finding the most outdated computer and smallest screen at hand and seeing how, or even whether, it is possible to complete a new study on it before releasing it to the general public. More generally, experimenters should try to work out what minimum criteria need to be met for participants to take part in the study and include those in recruitment instructions.

Another factor to consider, particularly when designing a study for research assistants to run, is what demands your study places on the researcher. For example, Case 2 requires the experimenter to advance through a series of scenes in OBS while interacting with a child through Zoom, and timing those interactions such that the interactions with the pre-recorded stimuli presented through OBS are convincing. It is certainly doable, but it does require some practice! It also imposes some demands on the experimenter's available hardware. We have found that older computers, particularly older Macs, have difficulty running both OBS and Zoom at the same time, and the video quality suffers heavily as a result. Particularly for studies that will be run by research assistants, it is important to ensure that those research assistants have access to adequate hardware to actually run the study. This is just one example, but in general, when designing an online study, researchers should consider how easy or difficult it will be for the experimenters to actually run with the required software.

One key design decision in terms of how difficult a study is for the experimenter is how the data are recorded. For

unmoderated studies, the data, particularly audio and video data, will inevitably have to be coded offline by researchers. This is also an option for many moderated studies, particularly if the study is challenging to run already. For example, another study in one of our labs using the same approach as Case 2 was found to require so much attention by the experimenter just to execute that we elected to code all the data off-line rather than trying to note participants' responses during the procedure. Some of the paradigms described above side-step this issue. For example, studies using PyHab, like Case 4, record data in the process of running the study with no additional effort. Of course, to ensure that the data are reliable, even in cases where the data are coded during the procedure, it is often worth having the data re-coded offline.

Data: Attrition and Quality

While many of the case studies described above are still in process, we have collected enough data to examine attrition and more generally whether the data are of comparable quality to in-lab data, and in some cases how well the data align with results acquired in the lab. Case 1, for example, was run partially in person and partially online, and we conducted a comparison of the data collected online to the in-person data and found no reliable differences (Kaminsky et al., 2021b). However, researchers cannot take for granted that this will be true for every study, and we have encountered different challenges in the different studies we have run.

Unmoderated studies face particular challenges, as noted above. When the experimenter is present, they can deal with obvious issues, for example 'is there actually a child participating in this study.' The study described in Case 6, in particular, ran into issues of non-serious participants (i.e., adults who took the study themselves just to get the participant compensation). As described above, when the project initially launched, the *majority* of participants were non-serious participants, and it was necessary to implement several types of screening to disincentivize these attempts to exploit the study for profit. Studies on Lookit, or other unmoderated studies that video-record participants, do not typically have this problem. A second issue is that the data itself is sometimes unusable for other reasons. For example, in the initial validation studies of Lookit, 35% of the videos recorded proved to be unusable due to some recordings failing for technical reasons, or the participant not being visible in the recording (Scott and Schulz, 2017). The technology has improved since then, but for any custom-designed solutions, extensive testing is needed to ensure that data are not lost due to technical issues, and piloting is strongly recommended to identify other potential problems in the data prior to opening the experiment to full data collection.

However, one note of caution for all online developmental research is that, relative to the history of our field, it is very, very new. We don't know how comparable online data are to in-lab data for many paradigms, and there are very few systematic comparisons across in-lab and online data with children (Scott and Schulz, 2017; for work with adults see, e.g., Weigold et al., 2013; Hauser and Schwarz, 2016). The pandemic has likely created a number of 'natural experiments' like Case 1 (Kaminsky

et al., 2021b), i.e., studies that started in person and moved online, that may provide further insight on the matter as they are published (indeed, we suspect other papers in this collection may do exactly that).

CLOSING THOUGHTS

In this paper we have attempted to provide a reference for researchers considering online developmental studies, to help them find the best tools and techniques for their particular needs. Broadly, by focusing on the key methodological constraints of a study, it can be relatively straightforward to identify the best tools for the job. There are some potential constraints we have not discussed in detail. Notably, the technical expertise available to the researcher can affect what solutions are actually achievable. Access to programming expertise, and particularly web development, can vastly expand the set of approaches available to a researcher, but these skills are not widely taught in our field. Universities sometimes offer such technological expertise to faculty in the form of dedicated research technology staff, but this is far from universal. However, most of the tools listed in this paper require no specific technical skills or programming ability, and were selected for this paper because they are accessible to researchers at any career stage and level of technical expertise. Furthermore, the majority of them are free, or at least inexpensive, and those that are not can often be licensed at the university level, making them affordable for individual researchers.

Evaluating the methodological constraints of a study and determining how to conduct it applies as much to in-lab research as it does to online research; in some cases the tools are even the same. PyHab (Kaminsky, 2019) and PsychoPy (Peirce et al., 2019) are both designed for in-lab and online studies, and there's no reason that any of the online presentation methods described here can't be employed in the lab as well, and with a little more methodological flexibility. For example, Case 1 was originally conducted in-person using a Qualtrics survey presented on a tablet, because the experimenter could control advancement through the survey when they were physically present to click the 'next' button. Of course, in-person research also opens up a host of additional methodological possibilities, like neuroimaging, pupillometry, and eye-tracking, that simply can't be done online with current tools. The methods that can be used for online research may also expand as new technology is developed: while it's unlikely that we'll ever be able to do remote fMRI, PET, MEG, or EEG studies, there is online eye-tracking for adult participants¹³, and developmental applications are currently under investigation, with some recent successes using offline analyses of videos to get fixation data (e.g., Chouinard et al., 2018; Chang et al., 2021).

¹³<http://turkergaze.cs.princeton.edu/>

Even when online studies are methodologically comparable to in-person data collection, they provide another source of participants that could perhaps represent a different population than is available in the lab, for better or worse. Thus, we come to a final point of consideration for all developmental research: participant recruitment. In-lab studies often recruit through databases of families that have expressed interest in participating in research, populated by purchased lists, state records, or other means. Of course, the families that actually participate are typically going to be those that are close to the lab itself, meaning that the demographics of a particular in-lab study will depend heavily on the lab's location and the local populace, or at least on experimenters being able to physically port the "lab" to other locations. Museum-based studies face a different set of potential constraints, in some cases recruiting from a more representative population and in other cases recruiting from a narrower population that have the resources, time, and interest in visiting such locations (Callanan, 2012). For online studies, geography and (e.g.) admission fees are no longer relevant restrictions, but having reliable high-speed internet access may restrict the population in ways that have not been fully quantified (for further discussion of these issues see Lourenco and Tasimi, 2020). Different populations may also be more or less accessible by different recruitment approaches (e.g., advertising on Facebook or Google versus recruiting for online studies from an existing database). One promising recent development is a centralized website (like ChildrenHelpingScience.com or LookIt) for developmental researchers to advertise their studies to families, which allows all the labs using the website to benefit from each others' recruitment practices, thereby potentially providing a much broader and more representative population than any one lab alone would be able to achieve (Sheskin et al., 2020). However, there are unavoidable minimum requirements for all of the online studies discussed here, and indeed nearly all online studies in principle: the participating family must have a device with internet access, a microphone, and in many cases a camera, that can be used in relative privacy, and researchers must take this limitation into account in interpreting their results.

REFERENCES

- Bass, I., Mahaffey, E., and Bonawitz, E. (2021). "Do you know what i know? children use informants' beliefs about their abilities to calibrate choices during pedagogy," in *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*, eds T. Fitch, C. Lamm, H. Leder, and K. Teßmar-Raible (Cognitive Science Society).
- Bechlivanidis, C., and Lagnado, D. A. (2016). Time reordered: causal perception guides the interpretation of temporal order. *Cognition* 146, 58–66. doi: 10.1016/j.cognition.2015.09.001
- Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., and Schulz, L. (2011). The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. *Cognition* 120, 322–330. doi: 10.1016/j.cognition.2010.10.001
- Bridges, D., Pitiot, A., MacAskill, M. R., and Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ* 8:e9414. doi: 10.7717/peerj.9414
- Callanan, M. A. (2012). Conducting cognitive developmental research in museums: theoretical issues and practical considerations. *J. Cogn. Dev.* 13, 137–151. doi: 10.1080/15248372.2012.666730
- Chang, Z., Di Martino, J. M., Aiello, R., Baker, J., Carpenter, K., Compton, S., et al. (2021). Computational methods to measure patterns of gaze in toddlers with autism spectrum disorder. *JAMA Pediatrics* 175, 827–836. doi: 10.1001/jamapediatrics.2021.0530
- Chouinard, B., Scott, K., and Cusack, R. (2018). Using automatic face analysis to score infant behaviour from video collected online. *Infant Behav. Dev.* 54, 1–12. doi: 10.1016/j.infbeh.2018.11.004
- Colombo, J., and Mitchell, D. W. (2009). Infant visual habituation. *Neurobiol. Learn. Mem.* 92, 225–234. doi: 10.1016/j.nlm.2008.06.002
- Hauser, D. J., and Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behav. Res. Methods* 48, 400–407. doi: 10.3758/s13428-015-0578-z
- Kominsky, J. F. (2019). PyHab: open-source real time infant gaze coding and stimulus presentation software. *Infant Behav. Dev.* 54, 114–119. doi: 10.1016/j.infbeh.2018.11.006
- Kominsky, J. F., Gerstenberg, T., Pelz, M., Sheskin, M., Singmann, H., Schulz, L., et al. (2021a). The trajectory of counterfactual reasoning in development. *Dev. Psychol.* 57, 253–268.
- Kominsky, J. F., Shafto, P., and Bonawitz, E. (2021b). "There's something inside": children's intuitions about animate agents. *PLoS One* 16:e0251081. doi: 10.1371/journal.pone.0251081
- Ultimately, once the COVID-19 pandemic has passed, the authors do expect to resume in-person data collection for many studies, but at the same time, we also expect to continue online data collection for others. For some designs, studies involving specific populations or specialized measures, in-person research will be preferable, but for others it may be easier or faster to continue to conduct research online. We believe this new wave of online developmental science will be long-lasting and bring many new benefits. Therefore, we expect that in the decades to come, developmental researchers will need to consider, for each new study, whether it is better to conduct it in person, online, or perhaps both at once, to make the most of all the methods that are now available to us.

AUTHOR CONTRIBUTIONS

JK wrote most of the original draft including two of the case studies, KB, IB, JC, and JL each wrote one of the case study sections. All authors participated in editing and revision.

FUNDING

Work described in this manuscript was funded in part by NSF SMA-1640816 to EB, a McDonnell Foundation Fellowship to EB, a Jacobs Foundation grant to EB and AM, and a UPenn MindCore postdoctoral fellowship to JL.

ACKNOWLEDGMENTS

The authors would like to thank all of their collaborators on the projects that are described in the case studies, including Elise Mahaffey, Patrick Shafto, Susan Carey, Hunter S. Liu, Skyler Courdrey, and Sophie Sharp, as well as members of the CoCoDev lab for thoughtful feedback on these methods, and many research assistants and participating families.

- Laloo, C., Pham, Q., Cafazzo, J., Stephenson, E., and Stinson, J. (2020). A ResearchKit app to deliver paediatric electronic consent: protocol of an observational study in adolescents with arthritis. *Contemporary Clin. Trials Commun.* 17:100525. doi: 10.1016/j.conctc.2020.100525
- Leonard, J., Lydon-Staley, D. M., Sharp, S. D. S., Liu, H. Z., Park, A., Bassett, D. S., et al. (2020). The toothbrushing task: a novel paradigm for studying daily fluctuations in young children's persistence. *PsyArXiv [preprint]* doi: 10.31234/osf.io/3hdur
- Liu, S. (2020). *Testing Babies Online Over Zoom*. Available online at: <https://medium.com/@shariliued/testing-babies-online-over-zoom-part-1-745e246b0af> (accessed August 31, 2021).
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Pearce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 51, 195–203. doi: 10.3758/s13428-018-01193-y
- Qualtrics (2005). *Qualtrics Online Survey Software*. Provo, UT: Qualtrics
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Scott, K., and Schulz, L. (2017). Lookit (Part 1): a new online platform for developmental research. *Open Mind*, 1, 4–14. doi: 10.1162/opmi_a_00002
- Sheskin, M., and Keil, F. (2018). TheChildLab.com a video chat platform for developmental research. *PsyArxiv [preprint]* doi: 10.31234/osf.io/rn7w5
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Weigold, A., Weigold, I. K., and Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychol. Methods* 18, 53–70. doi: 10.1037/a0031607

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Kominsky, Begus, Bass, Colantonio, Leonard, Mackey and Bonawitz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Natural Variability in Parent-Child Puzzle Play at Home

Nicole Pochinki¹, Dakota Reis¹, Marianella Casasola², Lisa M. Oakes³ and Vanessa LoBue^{1*}

¹Department of Psychology, Rutgers University, Newark, NJ, United States, ²Department of Psychology, Cornell University, Ithaca, NY, United States, ³Department of Psychology, University of California, Davis, Davis, CA, United States

Here, we observed 3- to 4-year-old children ($N=31$) and their parents playing with puzzles at home during a zoom session to provide insight into the variability of the kinds of puzzles children have in their home, and the variability in how children and their parents play with spatial toys. We observed a large amount of variability in both children and parents' behaviors, and in the puzzles they selected. Further, we found relations between parents' and children's behaviors. For example, parents provided more scaffolding behaviors for younger children and parents' persistence-focused language was related to more child attempts after failure. Altogether, the present work shows how using methods of observing children at a distance, we can gain insight into the environment in which they are developing. The results are discussed in terms of how variability in spatial toys and spatial play during naturalistic interactions can help us contextualize the conclusions we draw from lab-based studies.

Keywords: puzzles, spatial language, spatial skill, play, parent-child interactions

OPEN ACCESS

Edited by:

Yvette Renee Harris,
Miami University, United States

Reviewed by:

Sarah Pila,
Northwestern University, United States
Ruth Ford,
Anglia Ruskin University,
United Kingdom

*Correspondence:

Vanessa LoBue
vlobue@psychology.rutgers.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 June 2021

Accepted: 18 August 2021

Published: 16 September 2021

Citation:

Pochinki N, Reis D, Casasola M,
Oakes LM and LoBue V (2021)
Natural Variability in Parent-Child
Puzzle Play at Home.
Front. Psychol. 12:733895.
doi: 10.3389/fpsyg.2021.733895

INTRODUCTION

Spatial skills are central for everyday functioning, allowing us to encode the features, locations, and orientations of objects, as well as mentally manipulate this information. Spatial skills not only make it possible to interpret maps and diagrams, but also they are important predictors of later achievement across diverse STEM disciplines (Wai et al., 2009; Uttal and Cohen, 2012). For decades, research has documented a significant and robust relationship between spatial skills and mathematics performance over the course of development (Smith, 1964; Guay and McDaniel, 1977; Brown and Wheatley, 1989; Casey et al., 1995; Shea et al., 2001; Wai et al., 2009; Pyers et al., 2010; Cheng and Mix, 2014; Verdine et al., 2016, 2017). As a result, identifying factors that might influence the development of spatial skills in early childhood has received a great deal of attention in the literature.

For example, researchers have examined children's constructive play, or play with toys that involve the manipulation of objects in space, such as jigsaw puzzles, shapes, or construction blocks. A large body of research has reported a positive relationship between constructive play in childhood and both advanced concurrent spatial abilities (Connor and Serbin, 1977; Serbin and Connor, 1979; Caldera et al., 1999) and enhanced spatial skills later in development (Newcombe et al., 1983; Baenninger and Newcombe, 1989; Dearing et al., 2012; Levine et al., 2012; Nazareth et al., 2013; Jirout and Newcombe, 2015). Further, a handful of interventions studies have shown a causal relation between children experiences with constructive play, and

a subsequent increase in various spatial skills (Casey et al., 2008; Bower et al., 2020; Schröder et al., 2020).

Importantly, such constructive play often occurs during interactions with parents. Thus, parents' behavior during such play may be important for developing spatial abilities as well. For example, Levine et al. (2012) found that parents used more spatial language, including words describing the spatial properties of objects (e.g., "big," "little," "flat," and "edge") when their children were engaged with more challenging puzzles. This finding is important because children who hear more spatial language perform better on spatial tasks (e.g., Szechter and Liben, 2004; Dessalegn and Landau, 2008; Casasola et al., 2009, 2020). Thus, exposure to language is one possible mechanism for how play with parents shapes children's developing spatial abilities.

Parents may also support children's emerging spatial skills during constructive play by giving feedback, structuring the task, and modeling ways to problem solve during constructive play (Wood et al., 1976; Gauvain et al., 2002; Mulvaney et al., 2006; Ralph et al., 2020; Thomson et al., 2020). Children whose mothers provided more support or scaffolding during a spatial task performed better on a cognitive capability test that included measures of spatial ability (Mulvaney et al., 2006). Further, several studies have shown parents who better communicate task objectives and provide appropriate feedback have children who perform better on spatial tasks and tests of spatial concepts (Casey et al., 2014; Lombardi et al., 2017). Thus, scaffolding is another mechanism by which parents may influence children's spatial development during play.

Altogether, a large and growing literature suggests that several factors—including constructive play, exposure to spatial language, and parent scaffolding—may all play a role in shaping the development of children's spatial skills. Importantly, many of these studies have been conducted outside of the home, typically in a lab setting, with specific constructive play toys and tasks provided to parents and children. Although such experimental control allows us to derive conclusions based on standard conditions, the sole use of such assessments is limited, as children's behavior, along with parents' behavior with their children, might differ in the lab when compared to this behavior at home. Moreover, the constructive toys provided for a study in the lab may differ from those with which children typically play. Indeed, parents themselves have a great deal of control over what types of spatial toys they make available for their children, and they have many options to choose from. A simple google search for children's spatial toys produced over 5 million results, which can be narrowed down by the type of spatial toy in which a parent is interested, along with price, and the age and gender of their child. And there is evidence that the types of toys with which children play might bring about specific types of behaviors. In fact, researchers have even suggested that gender differences in spatial abilities might be attributable to differences in the toys parents select for girls versus boys (Todd et al., 2016; Coyle and Liben, 2020).

The COVID-19 pandemic has put a number of constraints on researchers' ability to collect data with children in the lab and in some ways, necessitated new approaches to study

development. Here, we show how we used a videoconference platform (Zoom) to study spatial play at home from a distance, along with the spatial and constructive toys that parents typically choose for their children. The existing studies that have examined children and their parents playing with toys in the home have focused on the relationship between the *frequency* of spatial play and parent support (Levine et al., 2012), or parent language and children's performance on spatial tasks (Mulvaney et al., 2006; Pruden et al., 2011; Polinsky et al., 2017; Ralph et al., 2020). Here, we asked a different question. Specifically, we sought to characterize the variability in various factors linked to spatial skills in children during their naturalistic play with the spatial toys they had at home. We explored variability in the types of puzzles families of 3- and 4-year-old children interact with in their homes, and the nature of those parent-child interactions during naturalistic play. We conducted the study over Zoom, and simply recorded parents and children as they played.

MATERIALS AND METHODS

Participants

Children between the ages of 3 and 4 years and their parents were recruited *via* a Rutgers University maintained database to participate in an online study investigating the development of spatial skills in children ages 3 and 4 years. Forty-two dyads participated in the study. Eleven were not included in our final sample due to either deviation from the protocol ($N=3$) or lack of puzzles at home ($N=8$). The final sample included 31 children (14 female, $M_{\text{age}}=44.6$ months, $SD=6.32$, $\text{Range}=35.8\text{--}55.3$ months) and their parents. All except for two parents presented as female. Families identified as White ($N=28$), Asian ($N=2$), or Mixed Race ($N=1$). Across all racial categories, four identified as Hispanic or Latino (three were White and one was Mixed Race). All caregivers had earned a bachelor's degree and 23 held advanced degrees. Our sample was middle class, with 22 families reporting an annual income above \$100,000, and only one family reporting an annual income below \$40,000. The Rutgers Institutional Review Board approved all procedures.

Procedure

Parents were invited to participate in an online study. Once an appointment was scheduled, families were emailed a link to a secure online survey *via* Qualtrics. This survey contained a consent form and an extensive questionnaire designed to describe the children's home playing environment. This questionnaire was part of a larger study designed to quantify the number and kinds of spatial toys in the participants' homes, and most of it will not be reported here. In one section of the survey, parents were presented with sample photographs of jigsaw puzzles and puzzle boards and were asked if they had those or similar toys at home. Parents were then asked to submit photos of those toys. The photos were used to code properties of the puzzles parents and children played with during our study.



FIGURE 1 | Camera set-up for the puzzle session.

One day prior to the study, participants received a reminder email informing them that they would be playing with puzzles. Parents were asked to select two puzzles from the ones they described in the survey for use during the study. The study itself was conducted on Zoom. On the day of the study, a researcher informed participants that they would be recorded playing with their child. Parents were asked to set up the camera in a high angle so all the pieces and playing space were in view and the researcher was able to look down at the participant's hands and all the pieces (see **Figure 1**). The researcher asked the parents to retrieve the previously selected puzzle(s). Parents and children were then instructed to play with each puzzle as they normally would for 10 min. If participants finished both puzzles before the 10-min mark, they were asked to retrieve additional puzzles. Thus, some children completed one puzzle during the 10-min session, while others completed up to 5. If they did not complete the puzzle during the 10-min session parents and children were given the option to finish. The researcher turned off her camera during the play period so that the parent and child could no longer see the researcher observing, and the researcher did not interrupt the play period before the 10-min mark.

Coding

Coders watched the recorded play session to categorize the puzzles' difficulty and to identify instances of specific child and parent behaviors. Children's insertion attempts, parental scaffolding behavior, and parental language were all coded using the open-source behavioral coding software, Datavyu.¹

Puzzle Difficulty

Parents chose puzzles that varied on a number of characteristics. One coder viewed all sessions and characterized all of the selected puzzles based on dimensions that might influence puzzle difficulty. There were five nested dimensions, each that were assigned a value of 0 (easiest) to 1 (most difficult). The first dimension was Puzzle type, which referred to whether the puzzle was a board puzzle (0) or jigsaw puzzle (1) (see **Figure 2A**). Puzzles were further coded for

whether or not they had a tray (0 if they did and 1 if they did not; **Figure 2B**). Puzzles that had a tray were then coded for whether they contained a background image that matched the puzzle piece (0) or no background image (1) (see **Figure 2C**). Puzzles that contained large pieces (i.e., pieces that were larger than the child's hands) were considered easier (0) than standard jigsaw puzzles (1) (see **Figure 2D**). Finally, puzzles were coded for whether or not they involved interlocking pieces (no interlocking=0 and interlocking=1; see **Figure 2E**). These dimensions were summed. For example, a jigsaw puzzle (1) with a tray (0) that contained a background image (0) with large (0) interlocking (1) pieces would receive a score of 2.

The number of pieces in each puzzle was also coded from the videos of the play session and from the puzzle photos submitted through the Qualtrics questionnaire. If information about the number of pieces was missing, an online search was conducted to identify the puzzle and obtain the specifications from the manufacturer's Web site. A second coder coded 25 puzzles out of a total of 65, and reliability was calculated for all the classifications described above ($\kappa=1$) and for the number of pieces (percent agreement=96%).

A puzzle difficulty composite score then was created by adding the binary values of all the coded difficulty dimensions and a code ranging from 1 to 5 based on the number of pieces such that the puzzles contained (i.e., 1 to 10 pieces received a score of 1; 11 to 20 pieces received a score of 2; 21 to 30 pieces received a score of 3; 30 to 40 pieces received a score of 4; and greater than 40 pieces received a score of 5). The final puzzle difficulty score ranged from 1 to 10, where a score of 10 was the most difficult.

Parent Behaviors

Two coders identified parent scaffolding events in the play session. Scaffolding events consisted of the sum of four different behaviors: (1) removing a piece that was placed in an incorrect space by the child, (2) helping by handing the child individual pieces or rotating pieces for the child, (3) pointing or outlining to a piece or a space in the puzzle, (4) pointing or outlining to the pictorial representation of the puzzle. Inter-rater reliability was calculated for piece removal ($\kappa=0.85$), helping ($\kappa=0.82$),

¹www.datavyu.org






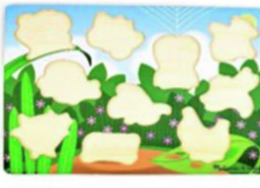


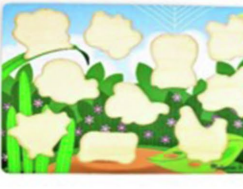

| | Easy (0) | Difficult (1) |
|---|---|--|
| A. Puzzle Type (Jigsaw vs. Puzzle Board) |  |  |
| B. Tray |  |  |
| C. Color-Matching Background Image |  |  |
| D. Piece Size |  |  |
| E. Interlocking |  |  |

FIGURE 2 | Puzzle dimensions that were coded for difficulty. **(A)** Puzzle type, **(B)** Tray, **(C)** Color-matching background image, **(D)** Piece size, **(E)** Interlocking.

pointing to ($\kappa=0.81$) or outlining ($\kappa=0.74$), a piece or space and pointing to/outlining a pictorial representation ($\kappa=0.92$). We created a total scaffolding score by summing the instances of each of these behaviors. In addition to scaffolding, we also coded instances where parents inserted a piece into the puzzle for the child ($\kappa=0.87$). This final code was not included in the total scaffolding behavior score.

Parental Language

One coder transcribed all parents' utterances. We defined utterances as vocalizations that were separated by grammatical closure, intonation contour, or prolonged pausing of more

than 2s. Three raters then coded each utterance to assess whether it contained spatial language (percent agreement=95%), praise (percent agreement=95%), or persistence-focused language (percent agreement=99%). Areas of disagreement were noted and resolved *via* discussion, ultimately resulting in consensus.

Spatial language was coded using a coding scheme developed by Cannon et al. (2007). Spatial language included any mention of spatial dimensions, shapes, locations and directions, orientations and transformations, spatial features, and properties. Examples of utterances coded as containing spatial language are "where's the flat edge?", "but I think you might need to rotate it a little," and "this is a big puzzle." Utterances that

contained more than one spatial word were not differentiated from those that contained only one spatial word. We only included spatial terms that were in reference to the construction of the puzzles and omitted terms that were unrelated to the puzzle (i.e., “Your blanket is under the bed”), or unrelated to its construction (i.e., “Put it in/on the puzzle”).

In addition to spatial language, which has been associated with children’s spatial ability in previous research, we also coded praise and persistence-focused language, which have been linked to more general engagement and persistence in children (Kelley et al., 2000). Praise was coded using a coding scheme developed by Gunderson et al. (2013) and included utterances that positively evaluated the child or the child’s actions (e.g., “You’re good at puzzles”; “good job”), or utterances that expressed general positive valence toward the child but not directed at any specific action (e.g., “Awesome!”; “Yay!”). Persistence-focused language was coded using a coding scheme developed by Lucca et al. (2019) and consisted of utterances that were focused on trying or repeated attempts to complete a goal-directed action. Frequently, this consisted of phrases that explicitly referred to acts of trying (e.g., “You’re trying so hard!”).

Child Behaviors

First, a trained coder watched the play sessions and identified children’s insertion attempts. An insertion attempt was defined as the first time the child took one puzzle piece and proceeded to either join it with one or more additional pieces or place it in an opening in a puzzle tray. An insertion attempt could be either *successful* if the child placed the piece in the correct space or *unsuccessful* if the child failed to insert the piece correctly and proceeded to place the piece back down on the floor or table. Each time the child attempted to insert the same piece in any opening or location was counted as a single event, which ended when the child either successfully inserted the piece or placed it down. A second researcher coded 25% of the participants and reliability was calculated for the event matching by both coders; reliability was calculated for both correct ($\kappa=0.88$) and incorrect insertions ($\kappa=0.76$).

After coding initial insertion attempts, a trained coder went back to each insertion attempt and counted the number of times the children unsuccessfully attempted to insert a single piece before either successfully inserting it or putting it down. An unsuccessful attempt was coded every time the child tried to insert the piece into a different place in the puzzle or in the same place but in a different orientation. A different orientation was defined as a rotation of the piece more than 90 degrees. A second researcher coded 25% of the insertion instances for each participant. Reliability was calculated for the number of insertion attempts ($\kappa=0.81$).

RESULTS

Data Analysis Plan

The main goals of this study were to describe the range of puzzles families selected for the play session, to examine parents’

naturalistic behavior with their children at home while playing with each puzzle, and to examine the relation between parent’s scaffolding and spatial language and children’s behavior with the puzzles. Upon initial visualization of the data, we observed a great deal of variability in all of the variables we measured. Thus, instead of running a large number of inferential statistics, we primarily provide descriptive data of both parents’ and children’s behaviors with the puzzles that they chose to interact with at home. Then, we normalized our measures by totaling the number of behaviors in each 1-minute interval, and then averaging across those intervals, and ran a correlation matrix on puzzle difficulty level, parenting variables (e.g., parent scaffolding, number of parental insertion attempts, parental spatial language, parental persistence-focused language, and parental praise), and child variables (e.g., age, children successful attempts, children overall attempts, and attempts after failure). Finally, we ran a set of simple gender comparisons across all of normalized data, given that gender differences in spatial abilities have been reported in previous research (Levine et al., 2005, 2016; Pruden et al., 2011).

Puzzle Difficulty

As mentioned above, the puzzles that participants typically played with in their homes varied widely, which is evident by the distribution of difficulty scores across puzzles (see **Figure 3**). The mean puzzle difficulty score was 6.56 ($SD=2.17$), and the difficulty scores spanned nearly the entire coded range from 2 to 10. Only five participants played with puzzles with a relatively low difficulty score that ranged between 2 and 4; the majority of participants played with puzzles that had a difficulty score in the middle of the range ($N=17$, between 5 and 7), and nine additional participants played with puzzles that were more difficult, ranging in score from 8 to 10.

Parent Language and Behaviors

The distribution of parents’ scaffolding, use of spatial language, and praise is in **Figure 4**. Two things are immediately clear. First, parents were highly variable, with some parents exhibiting high levels of these behaviors and other parents exhibiting low levels of these behaviors. It is possible that some of the variability in the number of behaviors may be due to variation in the length of the session. Although parents and children were encouraged to play for 10 min, some dyads played for less and others played for longer ($M=9.77$ min, $SD=1.6$ min, range 5.6 min–12.1 min). To examine whether the length of the session was related to the frequency of parent or child behaviors, we conducted a series of correlations. None of the relations between the duration of the session (in seconds) and parent or child behaviors were statistically significant ($p>0.05$). However, we normalized the data for all inferential statistics (see Section “Data Analysis Plan”).

Second, the distributions for the parent behaviors are very similar, with relatively low levels of the behaviors occurring more frequently than relatively high levels of the behaviors. Further, there is some evidence that the same parents were exhibiting relatively high or relatively low levels of some

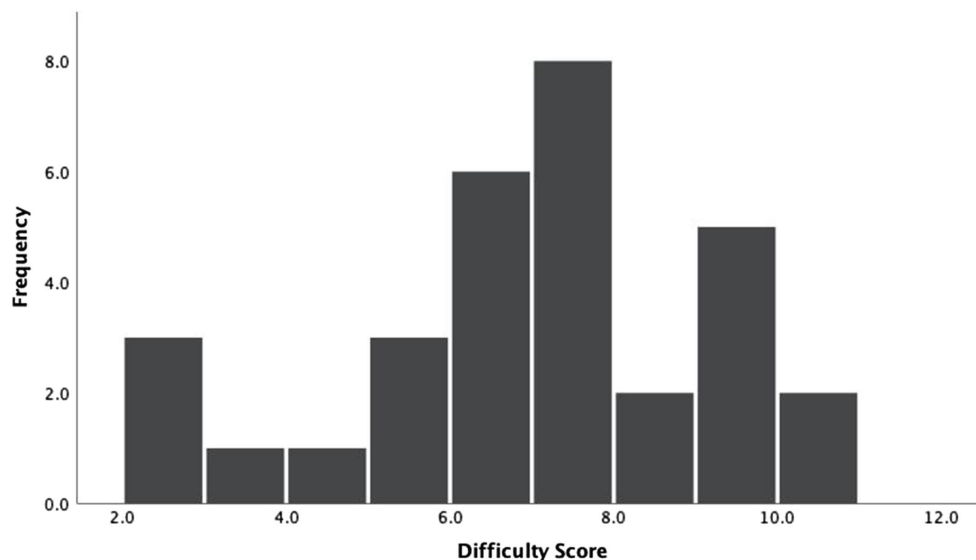


FIGURE 3 | Distribution of difficulty scores across puzzles.

combinations of these variables. For example, parent use of praise per minute was related to parent spatial language per minute, $r(31)=0.52$, $p<0.05$, and the relation between parent praise and parent use of persistence-focused language per minute was approaching significance, $r(31)=0.33$, $p=0.07$. This result suggests that there are effects of parental talk in general. Further, there were small, non-significant correlations between parent scaffolding behaviors and spatial language events per minute, $r(31)=0.28$, $p=0.13$, and praise, $r(31)=0.26$, $p=0.17$, suggesting that there were also parental behaviors specific to child behavior in this task.

Interestingly, utterances containing persistence-focused language were relatively rare, $M=1.61$ ($SD=1.61$), ranging from 0 to 5 across the session as a whole. Fifteen parents did not produce any utterances with this type of language at all.

To further understand parents' scaffolding behaviors, we examined separately the individual behaviors we coded. Recall that we coded parents' removal of an incorrectly placed piece, handing or rotating pieces, pointing or outlining puzzle space, and pointing or outlining pictorial representations of the puzzle. Parents more often pointed to or outlined the pieces or the puzzle ($M=24.29$, $SD=18.07$), than rotated or handed their child a puzzle piece ($M=9.97$, $SD=11.94$). Some parents simply inserted pieces into the correct places in the puzzle for the child, $M=5.06$ times per child ($SD=7.33$). There were large individual differences in this behavior; 21 parents rarely, if ever, inserted a piece for their child (ranging from 0 to 3 pieces), whereas 10 parents inserted between 7 and 30 pieces for their children.

Child Behaviors

Children's behaviors were also extremely variable. The distribution of total attempts and successful attempts to insert a piece is in **Figure 5**. In terms of attempts, children ranged from making

as few as 12 attempts to making as many as 81 attempts, suggesting individual differences in how interested children were in the puzzle. Children's successful insertions ranged from 1 to 41. The proportion of successful attempts ranged from 3 to 86%, again showing the extreme variability in children's behaviors.

We also coded how many times children attempted an insertion following a failed attempt. On average, children made 14.61 ($SD=9.5$), such attempts ranging from 2 to 37 attempts. Out of the 250 events where children tried to reinsert a piece upon failure, 66% had a successful outcome eventually.

Relations Among Variables

Next, we examined how parental behaviors were related to child behaviors during play. To account for the fact that participants' play time varied ($M=9.77$ min, $SD=1.6$, range 5.6 min–12.1 min), we normalized our measures by totaling the number of behaviors in each 1-minute interval, and then averaging across those intervals. Thus, our measures for these analyses were the number of behaviors or utterances per minute.

First, we examined how our measures were related to child age. The only relation between parental behaviors and child age was a negative correlation between age and parent scaffolding, $r(31)=-0.38$, $p<0.01$. Parents provided more scaffolding behaviors for younger children. It is also noteworthy that there was a small, non-significant relationship between age and children's successful attempts, $r(31)=0.28$, $p=0.12$, with older children demonstrating more successful attempts than younger children.

Interestingly, despite the wide variation in puzzle difficulty, we found that few child or adult behaviors were related to puzzle difficulty. There was no clear relation to child age, to parental scaffolding or language. The relation between puzzle difficulty and children's successful number of insertion

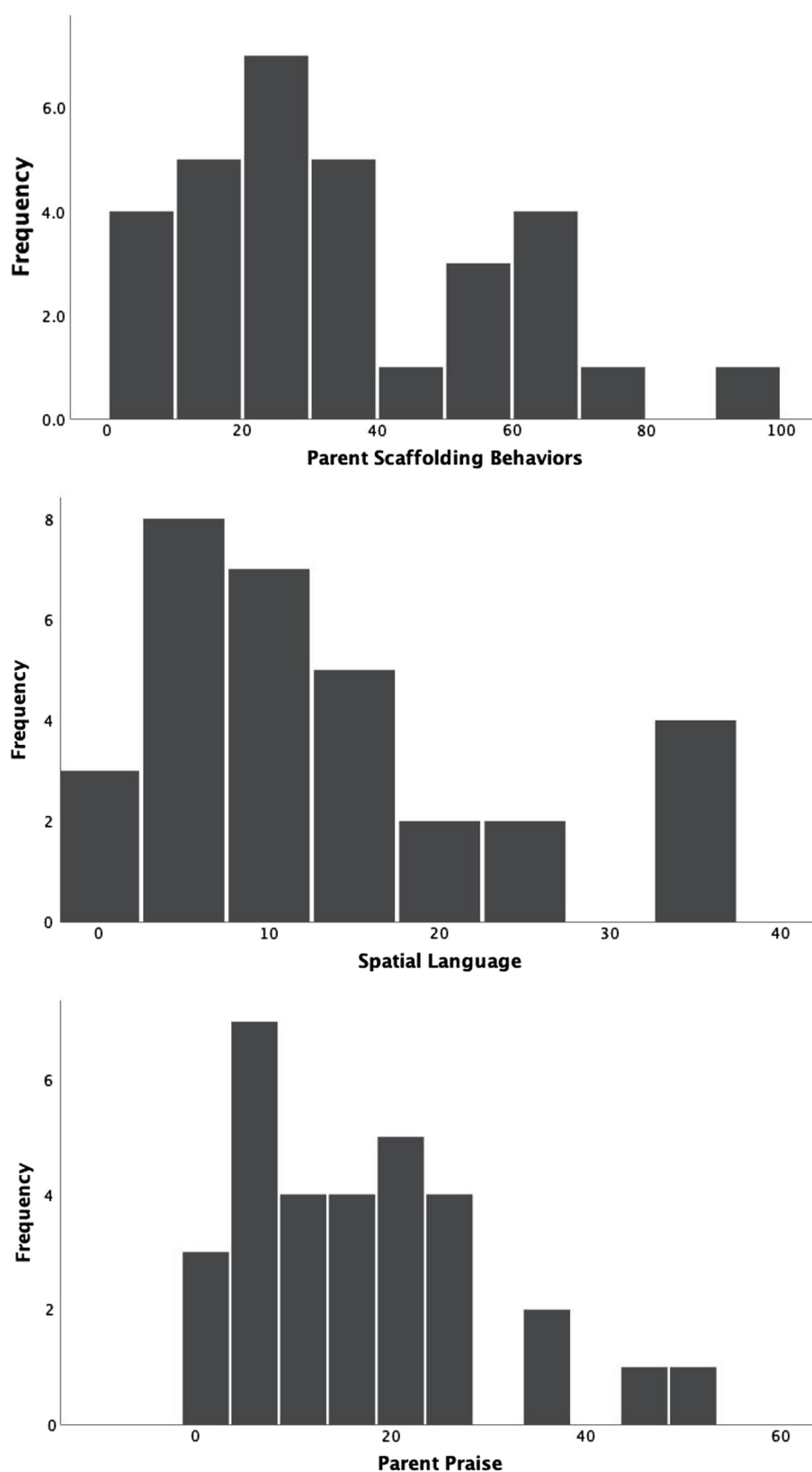


FIGURE 4 | Distribution of parent behaviors.

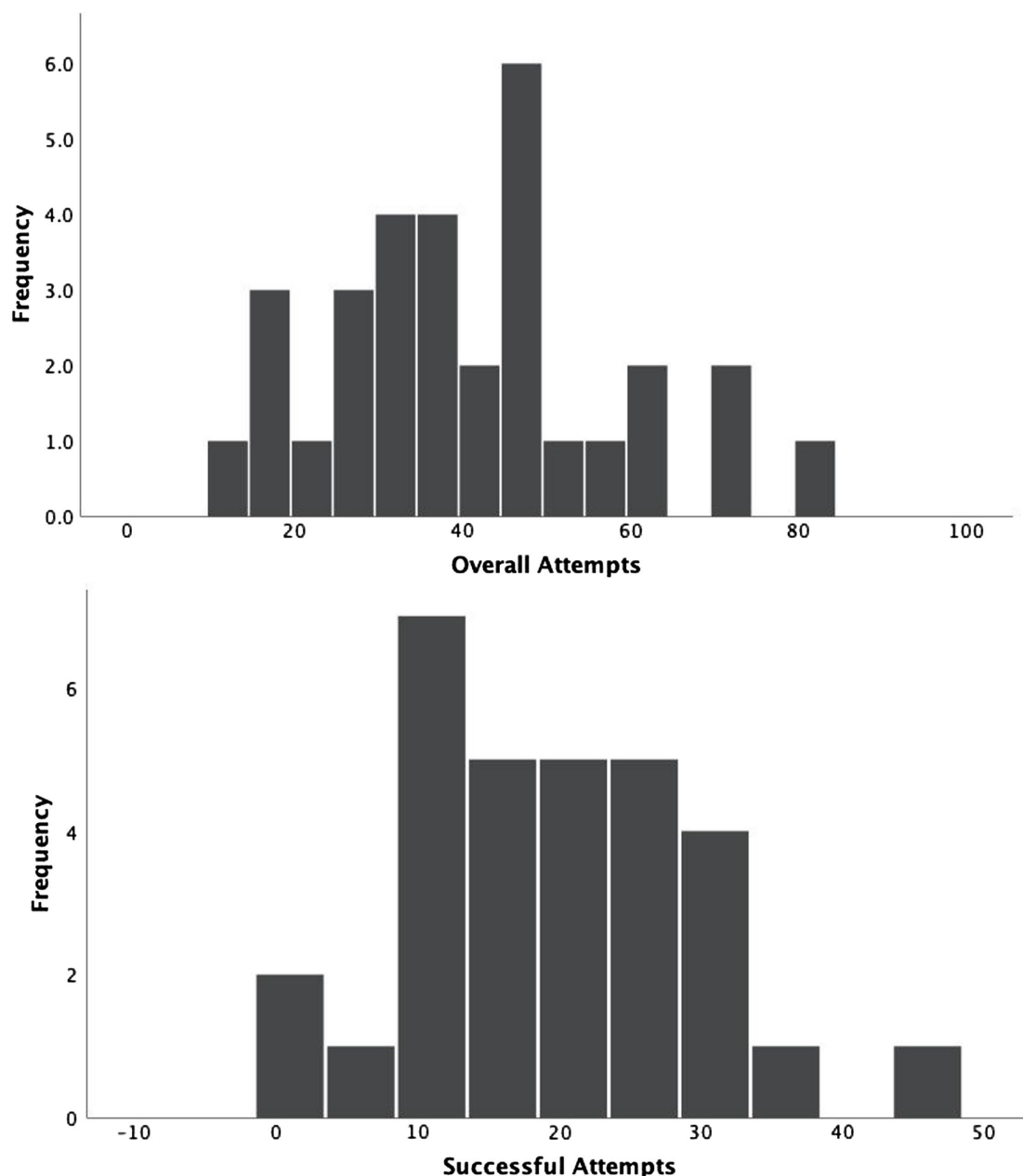


FIGURE 5 | Distribution of children's attempts.

events was approaching significance, $r(31) = -0.32$, $p = 0.08$. Not surprisingly, children were less likely to successfully insert a piece in more difficult puzzles. Note that we conducted a second set of correlations after removing the number of pieces from the difficulty score, as the number of pieces might have skewed the results. However, the results were the same.

We also found that parent and child's behaviors were related. In particular, the number of children's insertion attempts after failure was positively related to parents' persistence-focused language, $r(31) = 0.46$, $p < 0.01$, suggesting that children who

tried more after failing had parents that encouraged them to be persistent. In contrast, although non-significant, children's successful attempts were negatively correlated with all four parenting behaviors, suggesting that in general, children who had fewer successful attempts had parents who used more spatial language, praise, persistence-focused language, and scaffolding (see **Table 1**).

Gender Differences

Finally to evaluate any gender differences, we ran a series of *t*-tests comparing boys to girls on each of our measured

TABLE 1 | Correlation between variables.

| | Parent scaffolding | Parent praise | Parent persistence-focused language | Parent spatial utterances | Child successful insertion attempts | Child insertion attempts after failure | Child age | Difficulty score |
|--|--------------------|---------------|-------------------------------------|---------------------------|-------------------------------------|--|-----------|------------------|
| Parent scaffolding | 1 | | | | | | | |
| Parent praise | 0.256 | 1 | | | | | | |
| Parent persistence-focused language | 0.174 | 0.325 | 1 | | | | | |
| Parent spatial utterances | 0.281 | 0.516** | 0.100 | 1 | | | | |
| Child successful insertion attempts | -0.256 | 0.166 | -0.092 | -0.292 | 1 | | | |
| Child insertion attempts after failure | 0.049 | 0.185 | 0.463** | 0.152 | 0.246 | 1 | | |
| Child age | -0.376* | 0.188 | -0.272 | -0.088 | 0.282 | 0.016 | 1 | |
| Difficulty score | 0.042 | 0.040 | 0.131 | 0.106 | -0.320 | 0.085 | 0.147 | 1 |

*Correlation is significant at the 0.05 level (two tailed). **Correlation is significant at the 0.01 level (two tailed).

variables. There were no gender differences in terms of age (females $M=45.2$, $SD=1.7$; males $M=44.39$, $SD=1.49$) or difficulty of the puzzles (females $M=6.89$, $SD=2.11$; males $M=6.29$, $SD=2.24$). We did find a significant difference in the number of children's attempts after failure, $t(29)=2.19$, $p=0.021$, 95% CI $[-1.16, -0.04]$, with girls ($M=1.64$ attempts per minute, $SD=0.99$) attempting to place puzzle pieces more often after failure than boys ($M=1.04$ attempts per minute, $SD=0.51$). Thus, girls appeared to be more persistent than boys in their puzzle play. Further, we found that the difference in the amount of parents' persistence-focused language directed to boys and girls approached significance $t(29)=1.04$, $p=0.066$, 95% CI $[-0.15, 0.50]$, with parents using more persistence-focused utterances with girls ($M=0.13$, $SD=0.16$) than with boys ($M=0.07$, $SD=0.12$). None of the other parent or child variables differed as a function of child gender.

DISCUSSION

A large body of research has reported a positive relation between constructive play with toys like puzzles and developing spatial skills in children (e.g., Casey et al., 2008; Levine et al., 2012; Jirout and Newcombe, 2015; Bower et al., 2020). However, most of these studies were somewhat constrained, involving constructive play in a lab, and/or with a preselected and uniform set of constructive toys. Although the COVID-19 pandemic has kept many researchers away from the lab, it has offered us the opportunity to develop strategies for studying some of our basic research questions from a distance, by using tools like Zoom to examine what parents and children do in their own homes. Here, for the first time, we recorded parents and children interacting with puzzles of their choice at home and provided a descriptive account not only of their behaviors, but also of their behaviors in relation to the puzzles with which they most typically interact. Importantly, because we used Zoom, we may have observed more naturalistic behaviors

than if we had been present in the home with a video recorder and an experimenter in the room. The experimenter kept her camera off, and thus parents and children may have forgotten her presence.

The most noteworthy finding from this descriptive study is the enormous variability we observed in both children and parents' behaviors, and in the puzzles they selected for play. This study is the first of its kind in provide detailed characterization of the kinds of puzzles children have at their homes as well as the variability in parents' and children's behavior while engaging in home puzzle play. The puzzles themselves varied on a number of dimensions that we coded for difficulty. Some of the puzzles were typical jigsaw puzzles with interlocking pieces, while others were puzzle boards that had pieces with shapes that fit into specific places on a tray. Some of the puzzles had oversized pieces, presumably making them easier to place, while others even had a colorful background that matched the background of the puzzle pieces themselves, making it possible for children to use perceptual cues like color to match the pieces to their correct location. Some children played with puzzles that had less than 10 pieces, while other children played with 40 or 50 piece puzzles. No two play sessions were quite alike. These differences in the puzzles that children actually play with every day provide a context for studies of children's puzzle play that have used a narrow set of puzzles. Researchers often assume that findings from the lab uncover processes involved in children's puzzle play that reflect developmental changes in spatial ability. However, the variability in the types of puzzles available in children's homes has raised the possibility that participants in lab studies might differ substantially in their familiarity with the experimental stimuli.

Besides variability in the puzzles, there was also a great deal of variability in both parents' and children's behavior when interacting with the puzzles. There were a large number of parents who engaged in very few scaffolding behaviors, and very little spatial language, praise, and persistence-focused language during the parent-child interactions. Most parents

fell somewhere in the middle of the range, but there were also parents that produced an incredibly large amount of these behaviors, some with over 60 scaffolding behaviors in a 10-min play session, and upward of 30–40 praise and spatial language utterances. Further, parents who tended to use more spatial language also tended to use more praise and persistence-focused language, as evidenced by the significant correlations between these variables.

Children's behavior also varied widely, with some of our participants attempting to place pieces into the puzzles less than 10–20 times, alongside almost a third of our sample producing more than 50 attempts. Their accuracy varied just as widely: Most of the children placed less than 20 pieces correctly in the 10-min session, but some placed more than 30. Older children tended to place more pieces correctly than younger children.

Given this large amount of variability and our small sample size, it is unsurprising that we found few significant correlations between our variables. However, our results do suggest some basic patterns. Specifically, there were few relations with child age in our data, likely reflecting, in part, the relatively narrow age range we sampled. More surprising, despite the wide variation in puzzle difficulty, there was little relation between the level of puzzle difficulty and child age, child behavior, or parent behavior. Parents also showed some evidence of being sensitive to children's need for help. More persistence-focused language was related to more child attempts after failure. Interestingly, there was a hint that children's successful attempts were negatively correlated with all four parenting behaviors. If confirmed in a larger sample, this pattern would suggest that parents' language and scaffolding are related to children's success in puzzle play. Specifically, it is possible that parents recognized when children were having a difficult time and used more language and scaffolding to direct them. Likewise, it is also possible that parents' behavior impacted their children's behavior. Indeed, children who attempted to place more puzzle pieces after failure also tended to have parents who encouraged them more, thus it is possible that parents' persistence-focused language drove children to try harder.

Altogether, the variability we found in the puzzles themselves and in parent-child behaviors suggests that lab-based studies that impose a large number of constraints on children's behavior might not fully represent how children interact with spatial toys in their everyday environments. It is especially noteworthy that our sample was not particularly diverse. Indeed, most of our families were middle to high income, and even then, we had to eliminate eight families because they did not have two puzzles in their homes. While our sample was not ethnically and economically diverse and this limitation hinders our confidence to generalize our findings to a wider population, we expect that in a more diverse sample, we are likely to see considerably more variability than reported here. Lower income families, for example, might not have as many puzzles at home as middle to higher income families, and as a result, children's

behavior when engaging in spatial play might differ systematically by SES. Further, the puzzles we observed here, while variable, were all characteristic of toys in Western, industrialized countries. It is likely that the types of spatial toys available cross-culturally vary significantly, which could, in-turn, affect the types of spatial play in which children engage.

This is not to suggest that lab-based studies are not useful or important; indeed, they have provided the basis for even the current investigation. Indeed, imposing constraints on children's behavior allow us to narrow the focus of our research questions and ask more about the causal relations between variables. Further, it is important to acknowledge that the observational nature of this study was also limited in that the presence of the researcher, even with the camera off, may have changed parents' behavior in a way that is systematically different from completely naturalistic behavior. Nevertheless, this work highlights the enormous amount of variability that exists in children's spatial play at home in a very narrow sample, which has important implications for the conclusions we draw about lab-based studies that impose even more constraints on children's behavior.

It is also important to note that despite the large amount of variability reported here, there are some relationships documented in previous literature that were also evident in the current sample, speaking to their robustness. For example, similar to our results, several studies have shown that parents provide more assistance to younger versus older children during puzzle-building tasks (Wertsch et al., 1980; Casasola et al., 2017), suggesting that parents might adjust their behavior to fit different children's needs. Finally, we found several gender differences suggesting that girls were more persistent than boys, making more attempts to place pieces into the puzzle after failure, and that parents used more persistence-focused language with girls than with boys and gave girls more difficult puzzles. Gender differences in children's spatial ability and spatial play have also been reported in previous literature, usually attributing more advanced spatial skills to boys than girls, but these findings are controversial (Baenninger and Newcombe, 1989; Levine et al., 2012) and require further research.

CONCLUSION

In conclusion, despite its descriptive and non-causal nature, the current study informs us about the types of variability in spatial toys and spatial play we might expect in real-world settings and can help us contextualize the conclusions we draw from lab-based studies. Given the wide variability of puzzles available in children's homes, future research could examine how the different characteristics of puzzles determine the nature of the parent-child interactions and what aspects of these interactions support spatial skills development. Our study also suggests that more large-scale, naturalistic studies of children's spatial play in the home could be incredibly informative, providing us with important information about what types of

spatial toys best promote the development of spatial skills, and how the types of toys interact with both child and parent characteristics over time.

DATA AVAILABILITY STATEMENT

The datasets and coding manuals presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: LoBue, V., Pochinki, N., Oakes, L., and Casasola, M. (2021). Natural variability in parent-child puzzle play at home. *Databrary* available at: <https://nyu.databrary.org/volume/1334> (Accessed June 29, 2021).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Rutgers University Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

REFERENCES

- Baenninger, M., and Newcombe, N. (1989). The role of experience in spatial test performance: a meta-analysis. *Sex Roles* 20, 327–344. doi: 10.1007/BF00287729
- Bower, C., Zimmermann, L., Verdine, B., Toub, T. S., Islam, S., Foster, L., et al. (2020). Piecing together the role of a spatial assembly intervention in preschoolers' spatial and mathematics learning: Influences of gesture, spatial language, and socioeconomic status. *Dev. Psychol.* 56, 686–698. doi: 10.1037/dev0000899
- Brown, D. L., and Wheatley, G. H. (1989). "Relationship between spatial ability and mathematical knowledge," in *Proceedings of the 11th Annual Meeting of Psychology of Mathematics Education* (New Brunswick, NJ), 143–148.
- Caldera, Y. M., Culp, A. M., O'Brien, M., Truglio, R. T., Alvarez, M., and Huston, A. C. (1999). Children's play preferences, construction play with blocks, and visual-spatial skills: are they related? *Int. J. Behav. Dev.* 23, 855–872.
- Cannon, J., Levine, S., and Huttenlocher, J. (2007). A system for analyzing children and caregivers' language about space in structured and unstructured contexts. Spatial Intelligence and Learning Center (SILC) Technical Report.
- Casasola, M., Bhagwat, J., and Burke, A. S. (2009). Learning to form a spatial category of tight-fit relations: how experience with a label can give a boost. *Dev. Psychol.* 45, 711–723. doi: 10.1037/a0015475
- Casasola, M., Bhagwat, J., Doan, S. N., and Love, H. (2017). Getting some space: Infants' and caregivers' containment and support spatial constructions during play. *J. Exp. Child Psychol.* 159, 110–128. doi: 10.1016/j.jecp.2017.01.012
- Casasola, M., Wei, W. S., Suh, D. D., Donskoy, P., and Ransom, A. (2020). Children's exposure to spatial language promotes their spatial thinking. *J. Exp. Psychol. Gen.* 149, 1116–1136. doi: 10.1037/xge0000699
- Casey, B. M., Andrews, N., Schindler, H., Kersh, J. E., Samper, A., and Copley, J. (2008). The development of spatial skills through interventions involving block building activities. *Cogn. Instr.* 26, 269–309. doi: 10.1080/07370000802177177
- Casey, M. B., Beth Casey, M., Nuttall, R., Pizaris, E., and Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Dev. Psychol.* 31, 697–705. doi: 10.1037/0012-1649.31.4.697
- Casey, B. M., Dearing, E., Dulaney, A., Heyman, M., and Springer, R. (2014). Young girls' spatial and arithmetic performance: the mediating role of maternal supportive interactions during joint spatial problem solving. *Early Child. Res. Q.* 29, 636–648. doi: 10.1016/j.ecresq.2014.07.005

AUTHOR CONTRIBUTIONS

NP, VL, LO, and MC drafted the manuscript. NP collected the data. NP and DR oversaw coding. All authors designed the study and approved the final version of the manuscript.

FUNDING

This research and preparation of this manuscript were supported by a grant from the National Science Foundation (DS 1823489) to MC, LO, and VL. The funding agencies had no role in the design of the study or the collection, analysis, and interpretation of data or in writing the manuscript, apart from their financial contribution.

ACKNOWLEDGMENTS

We would like to thank all the families that participated in this study, along with Abigail Boatmun, Sharon Starling, Brenda Velarde, and Fatima Yassein for help with coding.

- Cheng, Y. L., and Mix, K. S. (2014). Spatial training improves children's mathematics ability. *J. Cogn. Dev.* 15, 2–11. doi: 10.1080/15248372.2012.725186
- Connor, J. M., and Serbin, L. A. (1977). Behaviorally based masculine- and feminine-activity-preference scales for preschoolers: correlates with other classroom behaviors and cognitive tests. *Child Dev.* 48, 1411–1416. doi: 10.2307/1128500
- Coyle, E. F., and Liben, L. S. (2020). Gendered packaging of a STEM toy influences children's play, mechanical learning, and mothers' play guidance. *Child Dev.* 91, 43–62. doi: 10.1111/cdev.13139
- Dearing, E., Casey, B. M., Ganley, C. M., Tillinger, M., Laski, E., and Montecillo, C. (2012). Young girls' arithmetic and spatial skills: the distal and proximal roles of family socioeconomics and home learning experiences. *Early Child. Res. Q.* 27, 458–470. doi: 10.1016/j.ecresq.2012.01.002
- Dessalegn, B., and Landau, B. (2008). More than meets the eye: the role of language in binding and maintaining feature conjunctions. *Psychol. Sci.* 19, 189–195. doi: 10.1111/j.1467-9280.2008.02066.x
- Gauvain, M., Fagot Craig Leve, B. I., and Kavanagh, K. (2002). Instruction by mothers and fathers during problem solving with their young children. *J. Fam. Psychol.* 16, 81–90. doi: 10.1037/0893-3200.16.1.81
- Guay, R. B., and McDaniel, E. D. (1977). The relationship between mathematics achievement and spatial abilities among elementary school children. *J. Res. Math. Educ.* 8, 211–215. doi: 10.2307/748522
- Gunderson, E. A., Gripshover, S. J., Romero, C., Dweck, C. S., Goldin-Meadow, S., and Levine, S. C. (2013). Parent praise to 1- to 3-year-olds predicts children's motivational frameworks 5 years later. *Child Dev.* 84, 1526–1541.
- Jirout, J. J., and Newcombe, N. S. (2015). Building blocks for developing spatial skills: evidence from a large, representative U.S. sample. *Psychol. Sci.* 26, 302–310. doi: 10.1177/0956797614563338
- Kelley, S. A., Brownell, C. A., and Campbell, S. B. (2000). Mastery motivation and self-evaluative affect in toddlers: longitudinal relations with maternal behavior. *Child Dev.* 71, 1061–1071. doi: 10.1111/1467-8624.00209
- Levine, S. C., Foley, A., Lourenco, S., Ehrlich, S., and Ratliff, K. (2016). Sex differences in spatial cognition: advancing the conversation. *Wiley Interdiscip. Rev. Cogn. Sci.* 7, 127–155. doi: 10.1002/wcs.1380
- Levine, S. C., Ratliff, K. R., Huttenlocher, J., and Cannon, J. (2012). Early puzzle play: a predictor of preschoolers' spatial transformation skill. *Dev. Psychol.* 48, 530–542. doi: 10.1037/a0025913
- Levine, S. C., Vasilyeva, M., Lourenco, S. F., Newcombe, N. S., and Huttenlocher, J. (2005). Socioeconomic status modifies the sex difference in spatial skill. *Psychol. Sci.* 16, 841–845. doi: 10.1111/j.1467-9280.2005.01623.x

- Lombardi, C. M. P., Casey, B. M., Thomson, D., Nguyen, H. N., and Dearing, E. (2017). Maternal support of young children's planning and spatial concept learning as predictors of later math (and reading) achievement. *Early Child. Res. Q.* 41, 114–125. doi: 10.1016/j.jecresq.2017.07.004
- Lucca, K., Horton, R., and Sommerville, J. A. (2019). Keep trying!: Parental language predicts infants' persistence. *Cognition* 193:104025. doi: 10.1016/j.cognition.2019.104025
- Mulvaney, M. K., McCartney, K., Bub, K. L., and Marshall, N. L. (2006). Determinants of dyadic scaffolding and cognitive outcomes in first graders. *Parenting Sci. Pract.* 6, 297–320. doi: 10.1207/s15327922par0604_2
- Nazareth, A., Herrera, A., and Pruden, S. M. (2013). Explaining sex differences in mental rotation: role of spatial activity experience. *Cogn. Process.* 14, 201–204. doi: 10.1007/s10339-013-0542-8
- Newcombe, N., Bandura, M. M., and Taylor, D. G. (1983). Sex differences in spatial ability and spatial activities. *Sex Roles* 9, 377–386. doi: 10.1007/BF00289672
- Polinsky, N., Perez, J., Grehl, M., and McCrink, K. (2017). Encouraging spatial talk: Using children's museums to bolster spatial reasoning. *Mind Brain Educ.* 11, 144–152. doi: 10.1111/mbe.12145
- Pruden, S. M., Levine, S. C., and Huttenlocher, J. (2011). Children's spatial thinking: does talk about the spatial world matter? *Dev. Sci.* 14, 1417–1430. doi: 10.1111/j.1467-7687.2011.01088.x
- Pyers, J. E., Shusterman, A., Senghas, A., Spelke, E. S., and Emmorey, K. (2010). Evidence from an emerging sign language reveals that language supports spatial cognition. *Proc. Natl. Acad. Sci. U. S. A.* 107, 12116–12120. doi: 10.1073/pnas.0914044107
- Ralph, Y. K., Berinhout, K., and Maguire, M. J. (2020). Gender differences in mothers' spatial language use and children's mental rotation abilities in preschool and kindergarten. *Dev. Sci.* 24:e13037. doi: 10.1111/desc.13037
- Schröder, E., Gredebäck, G., Gunnarsson, J., and Lindskog, M. (2020). Play enhances visual form perception in infancy—an active training study. *Dev. Sci.* 23:e12923. doi: 10.1111/desc.12923
- Serbin, L. A., and Connor, J. M. (1979). Sex-typing of children's play preferences and patterns of cognitive performance. *J. Genet. Psychol.* 134, 315–316. doi: 10.1080/00221325.1979.10534065
- Shea, D. L., Lubinski, D., and Benbow, C. P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: a 20-year longitudinal study. *J. Educ. Psychol.* 93, 604–614. doi: 10.1037/0022-0663.93.3.604
- Smith, I. M. (1964). *Spatial Ability: Its Educational and Social Significance*. San Diego: R.R. Knapp.
- Szechter, L. E., and Liben, L. S. (2004). Parental guidance in preschoolers' understanding of spatial-graphic representations. *Child Dev.* 75, 869–885. doi: 10.1111/j.1467-8624.2004.00711.x
- Thomson, D., Casey, B. M., Lombardi, C. M., and Nguyen, H. N. (2020). Quality of fathers' spatial concept support during block building predicts their daughters' early math skills – but not their sons'. *Early Child. Res. Q.* 50, 51–64. doi: 10.1016/j.jecresq.2018.07.008
- Todd, B. K., Barry, J. A., and Thommessen, S. A. O. (2016). Preferences for “gender-typed” toys in boys and girls aged 9 to 32 months. *Infant Child Dev.* 26:e1986. doi: 10.1002/icd.1986
- Uttal, D. H., and Cohen, C. A. (2012). Spatial thinking and STEM education: when, why, and how? *Psychol. Learn. Motiv.* 57, 147–181. doi: 10.1016/B978-0-12-394293-7.00004-2
- Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., and Newcombe, N. (2017). Links between spatial and mathematical skills across the preschool years. *Monogr. Soc. Res. Child Dev.* 82, 81–88. doi: 10.1111/mono.12284
- Verdine, B. N., Lucca, K. R., Golinkoff, R. M., Hirsh-Pasek, K., and Newcombe, N. S. (2016). The shape of things: the origin of young children's knowledge of the names and properties of geometric forms. *J. Cogn. Dev.* 17, 142–161. doi: 10.1080/15248372.2015.1016610
- Wai, J., Lubinski, D., and Benbow, C. P. (2009). Spatial ability for STEM domains: aligning over 50 years of cumulative psychological knowledge solidifies its importance. *J. Educ. Psychol.* 101, 817–835. doi: 10.1037/a0016127
- Wertsch, J. V., Gillian, D., McNamee, J. B., and McLane, N. A. (1980). The adult-child dyad as a problem-solving system. *Child Dev.* 51, 1215–1221. doi: 10.2307/1129563
- Wood, D., Bruner, J. S., and Ross, G. (1976). The role of tutoring in problem solving. *J. Child Psychol. Psychiatry* 17, 89–100. doi: 10.1111/j.1469-7610.1976.tb00381.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Pochinki, Reis, Casasola, Oakes and LoBue. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The Negative Impact of Noise on Adolescents' Executive Function: An Online Study in the Context of Home-Learning During a Pandemic

Brittney Chere and Natasha Kirkham*

Centre for Brain and Cognitive Development, Department of Psychological Sciences, Birkbeck, University of London, London, United Kingdom

OPEN ACCESS

Edited by:

Nicola K. Ferdinand,
University of Wuppertal, Germany

Reviewed by:

Julia Karbach,
University of Koblenz and Landau,
Germany
Stefania Muzi,
University of Genoa, Italy

*Correspondence:

Brittney Chere
bchere01@mail.bbk.ac.uk

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 26 May 2021

Accepted: 31 August 2021

Published: 22 September 2021

Citation:

Chere B and Kirkham N (2021)
The Negative Impact of Noise on
Adolescents' Executive Function: An
Online Study in the Context
of Home-Learning During
a Pandemic.
Front. Psychol. 12:715301.
doi: 10.3389/fpsyg.2021.715301

UNICEF estimates that 1.6 billion children across the world have had their education impacted by COVID-19 and have attempted to continue their learning at home. With ample evidence showing a negative impact of noise on academic achievement within schools, the current pre-registered study set out to determine what aspects of the home environment might be affecting these students. Adolescents aged 11–18 took part online, with 129 adolescents included after passing a headphone screening task. They filled out a sociodemographic questionnaire, followed by a home environment and noise questionnaire. Participants then completed three executive function tasks (the Flanker, the Backward Digit Span, and the Wisconsin Card Sorting Test) while listening to a soundtrack of either white noise or home-like environmental noise. For purposes of analysis, based on the noise questionnaire, participants were separated into quieter and noisier homes. Results revealed that measures of the home environment significantly correlated with individual perceptions of noise and task performance. In particular, adolescents coming from noisier homes were more likely to report that they studied in a noisy room and that they were annoyed by noise when studying. In terms of noise and task performance, the Flanker task revealed that while older adolescents were more efficient overall than their younger peers, those older adolescents from noisier homes seemed to lose this advantage. Additionally, reaction times for younger adolescents from noisier homes were less impacted by accuracy compared to their peers from quieter homes, though there was no difference for the older adolescents. This evidence suggests that higher in-home noise levels lead to higher rates of annoyance and may be hindering home-learning, with both younger and older adolescents being impacted. Furthermore, the long-term effect of in-home noise on adolescent executive function task performance indicates that these findings transcend the pandemic and would influence in-school learning. Limitations and advantages of online adolescent research without researcher supervision are discussed, including sociodemographics and adapting tasks.

Keywords: environmental noise, home-learning, adolescent development, COVID-19, online research, executive function

INTRODUCTION

Due to the COVID-19 pandemic, schools across 188 countries closed their doors to students by April 2020 in order to contain the spread of the virus, leaving approximately 1.6 billion students to continue their education from the safety of their homes¹. The impact this will have on the education of these children is vast and unprecedented. One particular question that needs to be addressed is how the change in environment, going from the structured classroom to the home, may be affecting educational outcomes. Secondary schools are often purpose built to foster learning, from the design and functionality of the entire building to the individual sections within classrooms (for an overview of U.K. regulations, see Department for Education and Skills, 2015). Importantly, the infiltration of noise from the outdoors and the transmission of noise between rooms within the building is often largely reduced. Even then, however, there is a large body of evidence showing that students' ability to learn is negatively affected by imposing noise, and that there may be possible long-term cognitive consequences (for an overview see Shield and Dockrell, 2003; Klatte et al., 2013). What, then, could this mean for learning within the home, an environment that is built to serve various functions and with the potential of having many and different distracting noise sources?

While noise pollution is presently government regulated, many researchers in the field would argue that stricter regulations need to be implemented and that more research is necessary based on the documented adverse effects from exposure to noise (Fink, 2017). Currently, the U.S. Environmental Protection Agency's general population guidelines are that the maximum average exposure to noise should not exceed 70 dB in order to prevent hearing loss, and that average indoor noise levels of 45 dB or greater will begin to interfere with activities and create annoyance². In a study attempting to determine the exposure to noise in schools, recordings across 13 U.K. schools during lessons revealed that the overall average noise level was found to be 64.2 dB L_{Aeq} , with a general background noise level of 51 dB L_{A90} (Shield et al., 2015). A study similarly attempting to establish the in-home noise levels of school children found that the average noise level in the main room of the home was 55.2 dB L_{Aeq} and the child's bedroom was 48.2 dB L_{Aeq} (Pujol et al., 2014). Evidently, similarly to schools, home noise levels seem to be exceeding the recommended indoor noise levels, making adolescents at risk for noise-induced annoyance and hindered learning.

As most formal learning occurs within the school environment, much of the research on how environmental noise impacts on learning takes place within schools. Not until the pandemic has the home been the environmental base for formal learning, with hardly any previous research, to our knowledge, having looked at the effect of in-home noise on learning. Thus, we will review the research focused on adolescent learning within schools, and link this to the few studies that have measured general in-home noise levels to determine how

these environments relate. The two streams of focus within the school literature are commonly the impact of noise on academic outcomes and the impact of noise on annoyance, with annoyance being defined as an emotional and cognitive response to a noise exposure (Guski et al., 2017) and is often used as an indicator for individual sensitivity to noise (Enmarker and Boman, 2004; Connolly et al., 2013). In a recent study by Massonnié et al. (2020) looking at the effect of noise on reported annoyance and schoolwork interference in children, they concluded that these are separate but correlated mechanisms that may be susceptible to individual differences. This last note concurs with Passchier-Vermeer and Passchier (2000), who determined that individual differences are key to understanding how noise affects development, particularly societal factors such as the home environment. Thus, while attempting to understand how in-home noise may be impacting learning using the research previously done within the school environment, we may further understand how adolescents' individual experiences within their home environments might be impacting their school learning, allowing for both streams of research to inform each other.

Unfortunately, most of the literature on noise annoyance and its effects on development are focused specifically on road traffic noise (Massonnié et al., 2020), giving a very narrow understanding of noise-induced annoyance as it is a subjective measure that can vary individually depending on the type of noise (Enmarker and Boman, 2004). What is known, is that levels of reported noise have been directly tied to noise-induced annoyance, with higher levels of noise relating to higher rates of annoyance in adolescents aged 13–15 (Ali, 2013) and 11–18 (Minichilli et al., 2018), though another study with 13-to-15-year-olds only found a poor correlation (Lundquist et al., 2000). Furthermore, adolescents have reported annoyance to both external and internal noises (Ali, 2013), though interestingly, one study found that adolescents aged 11–16 reported more annoyance for noise stemming from outside of the classroom compared to internal noise, even though noise from within the classroom occurred much more frequently (Connolly et al., 2013). The positive deduction made by the authors was that reducing the nuisance of outdoor noise heard within the classroom alone should then greatly decrease the negative effects of noise. Potential evidence for this was found by Ali (2013), who reported that noise levels within classrooms significantly reduced after restricting nearby outdoor road traffic and railways.

The findings looking at effects of age on rates of noise-induced annoyance are not as clear, with some studies showing that younger adolescents report more annoyance compared to their older peers (Lundquist et al., 2000; Ali, 2013; Minichilli et al., 2018), while another study found higher rates of noise-induced annoyance in older adolescents (Connolly et al., 2013). While effects of noise on academic performance within the literature of noise-induced annoyance is often not directly measured, students have reported the belief that environmental noise levels have negatively affected their academics (Ali, 2013; Connolly et al., 2013). To note, no effects of gender on rates of annoyance have been found (Lundquist et al., 2000; Enmarker and Boman, 2004; Minichilli et al., 2018). We can therefore conclude that levels of indoor classroom noise are directly tied to rates of annoyance and

¹ unicef.org

² epa.gov

self-reported academic performance. Such a finding is important, as Pujol et al. (2012) reported that indoor noise levels in the home have been linked to dwelling type, with children's bedroom noise levels being higher in more collective dwellings (i.e., being closer to and having more neighbors) and the main room in the home being higher in detached dwellings (i.e., no direct neighbors). Furthermore, they reported higher noise levels when more people were present in the home. This is particularly poignant in the context of COVID-19, whereby families may be grouping together in homes in order to support each other through the pandemic and the isolating lockdowns.

The line of research that has been directly measuring the effect of noise on academic achievement has largely focused on children (Connolly et al., 2019), with two in-depth reviews published. In the earlier review by Shield and Dockrell (2003), they determined that noise within the classroom appears to impact specifically on numeracy, reading, language, and speech, and also on overall academics. Furthermore, they deduced that noise is likely to have a greater impact when completing tasks that require higher processing demands, meaning that adolescents with more cognitively demanding schoolwork may be more affected by noise than the children included in the review. In Klatte et al.'s (2013) later review, they concluded that there is currently evidence for significant negative effects of environmental noise within the classroom on auditory tasks that involve the perception of speech and listening comprehension, as well as non-auditory tasks that involve reading, writing, and short-term memory. Of note, both reviews determined that due to the mixed findings in the literature, there is not enough evidence for a strong understanding and conclusion on how noise negatively impacts on academic learning. They do, however, state that chronic exposure to environmental noise is likely to impact on general cognitive development, meaning that exposure during childhood could have cascading effects on later adolescence and potentially adulthood. This would also imply that consistent exposure to noise in the home could also have widespread consequences.

A more recent study specifically looking at adolescents aged 11–16 found direct evidence of classroom noise affecting their reading ability. Connolly et al. (2019) had students listen to a naturalistic recording of non-verbal classroom noise through headphones at 50, 65, and 70 dB while completing a short reading task. By using an audio recording that depicts the actual environment that students typically learn in, along with a controlled learning paradigm, they were able to directly determine how this noise impacts on performance. Interestingly, they found that while older adolescents were generally better than their younger peers in the 50 dB condition, only the older adolescents' performance was negatively affected in the louder 65 dB condition. When comparing the audio at 50 and 70 dB, participants of all ages attempted less questions and were less accurate in the 70 dB condition. It was suggested that the greater effect of the 65 dB noise condition on the older adolescents was tied to their enhanced focus on the task, and they were thus more disrupted by the noise, but then at the highest noise level it was loud enough to be cognitively distracting for both age groups. It is therefore evident from much of the research looking into

how environmental noise impacts on learning, that the cognitive ability to deal with background noise plays a key role.

When environments are information rich with lots of stimulus input, learning requires the skills to attend to and process relevant and important information, while ignoring the irrelevant and distracting information (Stevens and Bavelier, 2012). As Parmentier (2014) concludes in a review, auditory distraction is defined as an auditory stimulus that violates what the cognitive system has predicted, therefore taking away attention from the task at hand and interfering with the current goal-directed behavior. Thus, the ability to selectively attend to the appropriate information and inhibit the distractors is necessary for learning to occur in most environments, giving executive function (EF) a fundamental role. Importantly, research shows that EF undergoes significant development throughout childhood and adolescence (Diamond, 2013). So while much of the research has historically focused on children, it is clear that adolescents are still very much susceptible to the negative effects of noise. Importantly, EF has also been directly linked to academic achievement across development (Jacob and Parkinson, 2015), particularly in math and reading (Best et al., 2011). More specifically, working memory, attentional flexibility, and inhibitory control have been found to be key EF components in this relationship (McClelland and Cameron, 2019). EF therefore serves two important roles within the current literature: (1) its developmental trajectory implies that differences within younger and older adolescents should be seen in terms of how noise impacts on learning, and (2) it is key for successful academic learning. The question remains though—how might all of this research translate to how noise within the home environment is impacting adolescent learning?

An important aspect to consider when making the comparison between the school and the home, is that adolescent formal learning usually involves teacher-guided group or independent learning, with other students of the same age generally working on the same tasks. These are factors that may indeed help keep focus and reduce the effects of noise and distraction. The environments at home, especially during the pandemic, most likely do not have any of these protective factors. Learning is now partially 'live' through an online format with stints of it being fully independent work without any teacher supervision (note: this will be dependent on the type of school and the school system). Additionally, very rarely would there be another person in the home working on the same task. Instead, with entire families being confined to the home during the pandemic, students learning at home could be surrounded by parents taking work calls, younger siblings playing, and grandparents making a cup of tea. Furthermore, the home is built to serve many functions other than studying and learning, ranging from cooking food, cleaning clothes, relaxing, playing, practicing hobbies, and being entertained. Thus, unlike a school, the home is not built to foster academic learning and to block out noisy distractors. So, while it is clear that noise has a negative impact on annoyance and academic achievement within schools, it is important to consider that the effect in the general home environment may be even greater. Of further importance is the finding that lower SES homes are likely to be exposed to higher levels of noise (Dale et al., 2015; Casey et al., 2017), and more chaotic homes have been

associated with lower family income and less educated caregivers (Dumas et al., 2005). This would imply that some adolescents may be more burdened by the impact of noise on their home-learning.

The main purpose of the current study was (a) to investigate the effect of the home environment, and in particular in-home noise, on adolescent learning in order to better understand how students have been impacted by the pandemic, and (b) to expand the methodology of online developmental research. To address these two aims, we ran an online experimental study along with questionnaires, where adolescents were asked to complete several EF tasks during which they listened to either a naturalistic recording of a noisy home or white noise. The EF tasks were used as a proxy for academic learning for two reasons: (1) the chosen tasks measure shifting, inhibitory control, and working memory, EF constructs that have been directly linked to academic achievement (McClelland and Cameron, 2019), and (2) the circumstances surrounding the COVID-19 pandemic made it impossible to work in close concert with schools and teachers to ascertain all of the learning materials being used by the individual participants, based on school year and age. Adolescents aged 11–18 were asked to take part with the plan of splitting them into younger and older age groups. An advantage of having adolescent participants is that it allowed for the study to be run independently online, without any researcher supervision. As Connolly et al. (2013) concluded that adolescents can both reliably and accurately report the noise acoustic levels within their environment and how it disrupts their learning, the current study had participants fill out a home environment and noise questionnaire. They were asked for specific details about their home and their subjective perceptions of the noise, which included measures of noise-induced annoyance. Based on their responses to the frequency of specific sound occurrences, they were then given an overall home noise score.

Participants then completed three different EF tasks: The Flanker, the Wisconsin Card Sorting Test (WCST), and the Backward Digit Span (BDS). The concurrent environmental noise being played through their headphones depicted a noisy and vibrant home, similar to what Connolly et al. (2019) had done with school noise, and following the previously mentioned review by Parmentier (2014), would represent the unpredictable and ever changing noise that is often present in a home and is most likely to cause distraction. The white noise was used to both serve as a constant background noise that was completely predictable and thus not distracting, and to block out actual environmental noise in the testing environment. The main experimental hypotheses were that overall, adolescents listening to the environmental noise would perform worse on all three EF tasks compared to their peers listening to the white noise. In terms of the effect of in-home noise on EF task performance, no specific hypotheses were predicted though these results were planned to be explored. It was, however, predicted that there would be an effect of experience with noise, whereby adolescents from quieter homes would have more difficulty on the tasks if listening to the environmental noise compared to their peers from noisier homes, who would have more practice in cognitively dealing with such distracting noise. Effects of age will be looked at, though exact predictions of the interactions with background

noise and in-home noise were not made due to the lack of previous studies directly measuring this.

MATERIALS AND METHODS

Participants

In total, 149 adolescents aged 11-to-18-years-old fully completed the online study from the comfort of their homes. As pre-registered, only the 129 who passed the headphone screening task (described below) who could ensure good audio quality were included. The mean age for these adolescents was 14.46 years ($range = 11.08$ – 18.92 years, $SD = 2.11$ years) with 74 females, 53 males, one gender fluid, and one not specified. To further ensure that the participants could appropriately see the visual stimuli presented on the screen and hear the auditory stimuli played through the headphones, they were asked about any visual or auditory impairments. Of these 129 participants, 101 reported no visual correction needed, 28 reported needing and wearing their corrective lenses, and none reported needing but not wearing their corrective lenses. Furthermore, 127 participants reported not needing a corrective hearing device, 2 reported needing and wearing their corrective hearing device, and none reported needing but not wearing their corrective hearing device. Thus, no participants were further excluded based on these criteria.

Data on participant ethnicity was collected and then grouped together based on the UK Office for National Statistics' ethnic groupings³. For detailed demographic information, see **Table 1**. Participants were recruited using flyers seeking neurotypical adolescents between the ages of 11 and 18-years-old in the United Kingdom ($N = 119$) and United States ($N = 10$) via word of mouth, social media, online parenting groups, and a database of participants. Data collection occurred between 16 June 2020 and 11 April 2021. For a very detailed accounting of the COVID-19 responses and regulations within both the United Kingdom

³www.ons.gov.uk

TABLE 1 | Subject ethnicity by income-to-needs ratio.

| Ethnicity | n (%) | INR Quartile Groups | | | |
|------------------------|------------|---------------------|----|----|----|
| | | 1 | 2 | 3 | 4 |
| Arab | 0 | 0 | 0 | 0 | 0 |
| Asian or Asian British | 6 (5.77) | 3 | 0 | 2 | 1 |
| Black or Black British | 3 (2.88) | 2 | 1 | 0 | 0 |
| Mixed | 10 (9.62) | 5 | 1 | 2 | 2 |
| White | 85 (81.73) | 16 | 35 | 12 | 22 |

INR stands for income-to-needs ratio. INR quartile groups were created using the quartile cut offs of 16,875, 21,875, and 25,000. Frequencies of INR quartiles per ethnicity are reported above. Participant ethnicity was grouped. Arab was its own group. Asian or Asian British = Indian, Pakistani, Bangladeshi, Chinese, and any other Asian background. Black or Black British = Caribbean, African, and any other Black background. Mixed = White and Black Caribbean, White and Black African, White and Asian, and White and any other Mixed background. White = British, Irish, and any other White background.

and United States, please see the following website⁴ describing the Oxford COVID-19 Government Response Tracker and its findings or see their published papers: United Kingdom-Cameron-Blake et al. (2020) and United States- Hale et al. (2020). Of note, all schools in both countries were closed for extended periods of time due to the pandemic, meaning that all participants in the current study experienced home-learning. Online written consent was obtained from each participant as well as from a caregiver, for those younger than 16 years of age. A £/\$5 gift voucher was given to each participant to thank them for their time. This study was designed in accordance with the Declaration of Helsinki and reviewed and approved by the School of Sciences Ethics Committee at Birkbeck, University of London, reference number: 192071. The analysis plan for this project was preregistered on asprecited.org on the 5th of August 2020, reference number: 45752. Prior to this date, no data from this project was accessed or analyzed.

Materials and Stimuli

The study was built and hosted on Gorilla Experiment Builder⁵. The executive function tasks were previously created on Gorilla to be used by experimenters. Participants completed the study via a link sent to them by the experimenter on a desktop computer or laptop device that was available to them. They were also asked to use any set of headphones that they had access to. Participants filled out two questionnaires and completed three executive function tasks.

Sociodemographic-Short Questionnaire

The MacArthur Research Network on SES and Health (2008) Sociodemographic-short questionnaire was used to measure several facets of socioeconomic status. Slight changes were made to reflect both American and British culture. The questionnaire included two visual ladders of sliding scales, measuring subjective perspectives of one's place within both the local community and the country. They were further asked about their highest level of education and their current job. In order to get an understanding of their income, they reported how much they earned in the past 12 months before deductions, how many adults bring income into the household, and how much total income they earned from all possible sources.

Home Environment and Noise Questionnaire

This three-part questionnaire, using a 4-point Likert response scale, was created for the purpose of this study (see **Supplementary Appendix A** for the full questionnaire). Part one asked the participant about the make-up of their household, including dwelling type and number of inhabitants before the presence of COVID-19 (before March 1st, 2020) and during COVID-19 (after March 1st, 2020). The second part asked questions regarding subjective noise measures, including their annoyance to the noise in their work rooms as well as desired

levels of noise for studying. The third section consisted of 25 questions asking about the frequency of specific noise sources in their homes. The questions themselves were designed so that half were positively stated, and the other half were negatively stated.

Noise Recordings

Two different audio recordings were played through the headphones during the completion of the EF tasks. Audacity 2.4.1⁶ was used to put together the two audio recordings that made up the Environmental Noise and White Noise conditions. Individual sounds within the environmental noise recording were obtained from Freesound⁷ and included the following: airplane, vacuum, toilet flush, footsteps, washing machine, muffled T.V. (words not interpretable), gaming laser sound, dog barking, door opening and closing, doorbell ringing, traffic, birds, various toys, and children laughing. The white noise (pure noise 3) was downloaded from The MC² Method online⁸. Both the White Noise and Environmental Noise recordings lasted for 15 minutes and were matched for frequency. A White Noise condition was used as a control to the Environmental Noise condition over silence as a means of blocking out the noises that would naturally be occurring in the participant's homes during the completion of the task and would thus bias results.

Headphone Screening

A headphone screening was used to (1) set the volume of the noise conditions, as we did not have direct control of the volume, and (2) to ensure the quality of the participant's headphones. The screening task was developed in Gorilla Experiment Builder by Brown et al. (2018). Participants pressed a 'play' button on the screen that played a white noise track. They were instructed to set the volume to the "loudest level that you can tolerate the sound without feeling like it's hurting your ears." After this, participants were played three sounds which were specifically developed to only be distinguishable through headphones (i.e., they could not be appropriately distinguished through the computer's speakers) and the participants were asked to determine if the first, second, or third sound was the quietest, as prompted on the screen. The correct answer was counterbalanced between being the first, second, and third tone played, with each repeated twice, giving a total of six trials. To pass the headphone screening, participants had to get five of the six trials correct. They moved onto the main tasks of the experiment once they passed and/or completed the three possible attempts. This allowed for all participants to have the chance to replace their headphones or to sort out any other issues before moving to the main tasks. The 19 participants that did not pass the headphone screening by the third attempt were not included in any analyses.

Flanker Task

This task was developed in the Gorilla Experiment Builder by Anwyl-Irvine et al. (2020) based on the original task by Rueda et al. (2004). The Flanker task is an attention network

⁴<https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker>

⁵www.gorilla.sc

⁶www.audacityteam.org

⁷www.freesound.org

⁸www.mc2method.org

test designed to measure inhibitory control. In the current experiment, participants were shown 5 arrows centrally on the screen. The middle arrow is referred to as the target arrow, with the surrounding arrows either appearing congruently and matching the direction of the central arrow or appearing incongruently and facing the opposite direction of the target arrow. The participants had to press the letter “z” on the keyboard if the target arrow was pointing to the left, or “m” if it was pointing to the right. They were asked to respond as quickly and as accurately as possible. The task began with 12 practice trials and feedback was provided for each trial informing them if they were correct or incorrect. The main task consisted of a total of 96 trials, which were separated into four blocks with a break in between. The participant had to press the spacebar to indicate that they were ready to begin the next block. For each trial, the arrows remained on the screen until the participant made a response. A central fixation cross appeared in between each trial with varying lengths of time (400, 600, 800, or 1000 ms). The task was counterbalanced in terms of the appearance of the central arrow (left or right) and the congruence of the surrounding arrows (congruent or incongruent). The trials and timings of the fixation cross were then randomized across participants. Dependent measures were based on reaction time (RT) and accuracy and are detailed in the results section.

Wisconsin Card Sorting Test Task

This version of the WCST was attained and further developed from Gorilla's task Samples⁹. This task was designed to measure set-switching and set-maintenance, utilizing abilities such as shifting, working memory, and inhibition (Huizinga and van der Molen, 2007). Each trial consisted of participants being given a target card to match with one of four other cards based on one of three rules: number (1, 2, 3, or 4), color (red, blue, green, or beige), or shape (circle, diamond, star, or triangle). This meant that three of the four cards to select from would each pair with the target card based on one rule alone, with the fourth card being a random card that did not match the target card. The task was designed to have 10 trials per rule block, with each rule repeating twice, giving a total of 60 trials. While the participants were aware of the three different rules, they were not told which rule to use nor when it would change. Therefore, they were only able to determine rule switches based on the 700 ms feedback they received after each trial. The rule block order for each participant was number, shape, color, shape, number, and color. The cards remained on the screen until the participant gave a response. There were a total of 64 unique cards which were pseudo-randomly displayed to ensure that there was no repetition of the target card, and that the different cards were spread out as non-target cards throughout and between the blocks. Dependent measures were based on errors made both within and between sets.

⁹<https://gorilla.sc/support/samples>

Backward Digit Span Task

This task was created using the Gorilla Experiment Builder¹⁰ by Massonnié (2020), though minor adjustments were made to add our two auditory conditions. The digit span task is commonly used to study memory, with arguments made that the forward digit span task more specifically measures short-term memory while the BDS task measures working memory (Wells et al., 2018). Participants were shown a series of numbers and were asked to respond by inputting the same numbers in reverse order. The first level began with two numbers, with each new level increasing by one additional number. Each level contained five trials whereby the participant needed to get three of the five trials correct in order to advance to the next level. This meant that three mistakes within a level led to the termination of the task. Each trial began with a 450 ms fixation cross followed by each number presented one at a time on the screen for 1500 ms, with 500 ms intervals. The numbers were displayed in pseudo random order whereby each number was random other than that the same number could not directly follow the previously displayed number. Participants were first given two practice trials with feedback on their performance to help ensure that they understood the task. Dependent measures were the total number of correct trials (final score) and proportion of correct trials throughout the task.

Procedure

Participants were first given information about the online study and, upon giving consent to participate, were then directed to begin. The study began by asking for the participant's age and gender. To monitor the study and any potential issues, participants were also asked if any of their siblings had taken part and their age, and if they themselves had previously attempted to participate in the study but did not complete it. They were then asked to specify if their previous lack of completion was due to loading delays/poor connection, needing to stop for time reasons, or to state some other reason.

The parent was then instructed to complete the Sociodemographic questionnaire, followed by both the parent and the adolescent completing the Home Environment and Noise questionnaire. The adolescent was then asked to put their headphones on and to complete the headphone screening task, whereby they had three chances to pass, although all participants continued to the experimental portion of the study regardless of passing or failing. Upon completing the headphone screening, participants went on to complete three tasks: (1) the Flanker task, (2) the WCST task, and (3) the BDS task. Participants were randomly assigned to either the White Noise or the Environmental Noise condition. If in the Environmental Noise condition, they completed all three tasks while listening to an audio recording simulating a ‘noisy home environment,’ while if in the White Noise condition, they simply listened to an audio recording of white noise. The order of the tasks was randomized for each participant and the exact same audio recording, depending on noise condition, began playing at the beginning of each task and stopped once the task was completed,

¹⁰<https://gorilla.sc/openmaterials/36699>

with the audio recording restarting each time. Therefore, they heard the same audio recording three times but for different lengths of time depending on the timing to complete each task.

Once the three tasks were completed, participants were then asked to state if the audio recording consistently played for the duration of each task, or if for any of the tasks the audio recording ended before the task was finished. They were then presented with a debrief of the study and were told that they had finished and could exit the browser window. The study took no more than 30 min to complete.

RESULTS

Scoring and Preprocessing Sociodemographic Questionnaire

Due to issues with collecting SES data through the study's online format, which is explained in detail in the discussion, only total family income was looked at. As overall total income is not very informative when considering the complexity of socioeconomic status, it was therefore decided to report families' income-to-needs ratio (INR). Total income was collected in bins, and INR was calculated by using the median of each income bin, similarly to King et al. (2020), and then dividing this number by the reported total number of inhabitants in the home before the pandemic. Calculated INRs were then grouped into quartile bins, with the break-down of ethnicity by INR quartiles seen in **Table 1**. While the current sample is perhaps slightly more heterogeneous than that often found within in-lab testing, it is still very much within the W.E.I.R.D. population. In terms of looking at income within later analyses, actual total income was used as income and number of inhabitants were individually investigated, and thus the combined INR measure was not used.

Home Environment and Noise Questionnaire

Responses to the 4-point Likert scale were added up to create an overall home noise score, where negatively phrased questions were reverse scored. The higher the overall score, the noisier the home was determined to be (lowest possible score = 25, highest possible score = 100). For part of the analyses, participants were grouped into noisier and quieter homes based on a median split ($Mdn = 64.5$).

Flanker

Two scores were pre-registered for this task. The Inverse Efficiency Score (IES) was developed to measure the participant's ability to efficiently complete the task in terms of both timing and correct responses ($IES = \text{mean reaction time/proportion of correct trials}$), with higher scores meaning less efficiency. Following Imburgio et al. (2020), the mean reaction time to the incorrect trials was subtracted from the mean RT from the correct trials to get a ΔRT Accuracy score. A higher positive score indicates a bigger difference between the two trial types, with an average longer RT on correct trials and an average shorter RT on incorrect trials. The opposite direction for the correct and incorrect trials led to higher negative scores. Lower scores closer to zero infer that the reaction times to correct and incorrect

trials are closer together and accuracy did not affect reaction time behavior. The congruency effect, here termed ΔRT Congruence, was further looked at as a measure of selective attention, whereby the mean reaction time on the incongruent trials was subtracted from the mean RT on congruent trials (van Leeuwen et al., 2007). Importantly, using these two difference scores allows for a better understanding of the effect of both accuracy and incongruency on performance, and removes the potential of simply looking at the effect of slow responders (Mullane et al., 2009).

As planned, four participants who reported audio issues during the task had their data removed. Furthermore, nine participants were excluded who did not pass the training (passing set at 8 out of 12 trials correct) and two whose performance was at chance level. All trials that were either less than 300 ms or greater than 1500 ms were removed (6.64% of total trials), and four participants with more than 25% of their data missing due to this criteria were excluded (van Leeuwen et al., 2007). After removing another two due to a combination of these issues, a total of 16.15% of the participants were removed from data analyses.

Wisconsin Card Sorting Test

As pre-registered, the WCST was scored based on errors made by the participant. Perseverative errors are those made based on following the rule from the previous set (does not apply to errors made in the first block), while non-perseverative errors are all other errors made. Of note, the first error made after a rule change is not counted as a perseverative error but as a non-perseverative error, as this is the first instance that the participant learns that the rule has changed. Any error after this that is made based on the previous rule set would then count as a perseverative error. The last score was failure to maintain set, established as the participant making an error after having gotten at least five correct in a row, all within the same rule block. The last score included was total errors made. Importantly, a single error made could be allocated toward one or more of the scores. For an overview of WCST scoring, see Cianchetti et al. (2007).

As planned, trials with a response time that exceeded 10s were removed (0.016% of all trials) (Piper et al., 2012). Although not planned, the nature of the task meant that those with worse internet connections experienced severe loadings delays. Additionally, several participants had long gaps in between trials. As both issues would strongly interfere with the participant being able to follow the rule sets, it was objectively determined to remove the data from four participants who took longer than two standard deviations above the mean to complete the task ($M = 3.27$ min, $2 SD = 6.65$ min), of which three of these participants experienced loading delays. No participants reported any audio issues, and only a total of 3.08% of participants were excluded.

Backward Digit Span

Both scores used were pre-registered. Final score is a commonly used measure (e.g., Lipsey et al., 2017) and represents the total number of correct trials. As participants could have achieved the same final level with either two errors per level or with none until the final level, we further looked at the proportion of correct trials to account for this difference.

Data from three participants were excluded as they did not follow the rules of the task (reported the numbers in forward order). As planned, data was excluded for those with audio issues, with 13 participants who had audio issues during the task and seven who self-reported having audio issues. A further three participants were excluded due to a combination of these issues. In total, 20.00% of participants were excluded.

Analyses

Home Environment, Subjective Noise Measures, and Executive Function

As pre-registered, analyses were performed to capture an understanding of the home environment, and how it may be affecting adolescents. As can be seen in **Table 2**, the number of inhabitants in the home during the pandemic both increased and decreased compared to the number of inhabitants before the pandemic hit. A paired-samples *t*-test revealed a small, though significant overall increase from the number of inhabitants occupying the home before the pandemic ($M = 3.97$, $SD = 1.11$) to during the pandemic ($M = 4.10$, $SD = 1.22$), $t(127) = -2.79$, $p = 0.006$. Although there was a significant difference in the number of inhabitants before and during the pandemic, only the number of inhabitants during the pandemic was used

in the following correlations as this number would be more representative of the adolescents' home environment when answering the questionnaires. Furthermore, though our sample was skewed toward participants from the United Kingdom, country of residence did not significantly correlate with any of the home or subjective noise measures, meaning that our sample did not significantly differ in the recorded home measures nor the subject noise measures across country of residence. Spearman bi-variate two-tailed correlations were run looking at the home environment and subjective noise measures (see **Table 3**). Correlations were Bonferroni corrected and significance was established at 0.00625 (0.05/8).

Of note, age did not significantly correlate with any of the home measures nor the subjective noise measures, meaning that younger adolescents were not more sensitive to noise, did not perceive more noise, nor were they more annoyed by noise than their older peers. Those who reported being more annoyed by the noise in the room they study in were significantly more likely to report higher annoyance to noise compared to their peers, and to be studying in a noisier room. Furthermore, number of inhabitants was also found to significantly correlate with dwelling type, with more inhabitants in the home being more likely to live in less collective dwellings. Higher home noise scores significantly correlated with more collective dwelling types and correlated with adolescents reporting more noise in the room they study in.

Interestingly, the correlations further revealed that those from noisier homes were more likely to report a preference for more background noise when studying while also being more annoyed by the noise in the room they study in; however, noise preference and noise annoyance while studying did not correlate. To further understand this seemingly contradictory finding, further analyses were done to determine if perhaps those from noisier homes are either likely to develop a preference to noise based on their exposure to it, or to become more annoyed by it. After grouping participants into noisier and quieter homes, however, those from noisier homes did not show the expected negative correlation, meaning that those who reported a preference for noise did not also report less noise annoyance, and vice versa.

Further spearman correlations were run to determine the relationship of task performance with both home measures and subjective noise measures (see **Table 4**). Correlations were Bonferroni corrected for multiple comparisons, with home measures being significant at 0.0166 (0.05/3) and subjective noise measures being significant at 0.0125 (0.05/4). Again, country of residence did not correlate with EF task performance. Results revealed that measures of the home did significantly relate to task performance. More inhabitants in the home during the pandemic significantly related to more perseverative errors and total errors on the WCST, and nearly significantly related to a higher Flanker Δ RT Accuracy score. Interestingly, being in more of a collective dwelling significantly correlated with a lower BDS final score but near significantly correlated with less total errors on the WCST. A more collective dwelling also nearly correlated with a better Flanker IES and significantly correlated with a lower Flanker Δ RT Accuracy score. As for the relationship between

TABLE 2 | Frequency of the type of home and the number of inhabitants in the home before and during the pandemic.

| | <i>n</i> | % | % Change | Mean |
|-------------------------------------|----------|-------|----------|------|
| Type of home | 128 | | | |
| Detached | 35 | 27.34 | | |
| Collective dwellings | 93 | 72.66 | | |
| Semi-detached | 24 | 18.75 | | |
| Terraced | 33 | 25.78 | | |
| Flat | 36 | 28.13 | | |
| Number of inhabitants before | 127 | | | 3.99 |
| 2 | 3 | 2.36 | | |
| 3 | 38 | 29.92 | | |
| 4 | 57 | 44.88 | | |
| 5 | 21 | 16.54 | | |
| 6 | 6 | 4.72 | | |
| 7 | 0 | 0.00 | | |
| 8 | 1 | 0.79 | | |
| 9 | 0 | 0.00 | | |
| 10 | 1 | 0.79 | | |
| Number of inhabitants during | 127 | | | 4.13 |
| 2 | 3 | 2.36 | 0.00 | |
| 3 | 35 | 27.56 | -2.36 | |
| 4 | 54 | 42.52 | -2.36 | |
| 5 | 21 | 16.54 | 0.00 | |
| 6 | 11 | 8.66 | 3.94 | |
| 7 | 0 | 0.00 | 0.00 | |
| 8 | 2 | 1.57 | 0.79 | |
| 9 | 0 | 0.00 | 0.00 | |
| 10 | 1 | 0.79 | 0.00 | |

TABLE 3 | Correlations between participant age, home measures, and subjective noise measures.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---------------------------------|---|--------|-------|--------|----------------|--------|---------------|---------------|---------------|
| (1) Age | 1 | −0.223 | 0.207 | −0.135 | 0.026 | 0.07 | 0.023 | 0.013 | 0.167 |
| (2) Total income | | 1 | −0.2 | 0.058 | −0.164 | 0.057 | 0.042 | −0.218 | −0.072 |
| (3) Home noise score | | | 1 | 0.051 | 0.243* | 0.123 | 0.261* | 0.348* | 0.240* |
| (4) Number of inhabitants | | | | 1 | −0.291* | 0.174 | −0.142 | −0.11 | 0.143 |
| (5) Dwelling type | | | | | 1 | −0.029 | −0.043 | 0.18 | 0.019 |
| (6) Comparative noise annoyance | | | | | | 1 | −0.106 | 0.05 | 0.309* |
| (7) Studying noise preference | | | | | | | 1 | 0.095 | 0.094 |
| (8) Room noise level | | | | | | | | 1 | 0.351* |
| (9) Room noise annoyance | | | | | | | | | 1 |

Age and home noise score were entered as continuous variables, all other variables are ordinal. $N = 106–128$

* $p < 0.00625$.

TABLE 4 | Correlations of task scores with home measures and subjective measures of noise.

| | Home measures | | | Subjective noise measures | | | |
|------------------------------------|---------------|--------------------------|---------------------------|-----------------------------|---------------------------|------------------|----------------------|
| | Total income | Number of inhabitants | Dwelling type | Comparative noise annoyance | Studying noise preference | Room noise level | Room noise annoyance |
| (1) Flanker IES | 0.081 | −0.011 | −0.215[†] | 0.047 | −0.032 | 0.070 | −0.035 |
| (2) Flanker Δ RT accuracy | 0.065 | 0.223[†] | −0.264* | 0.008 | −0.178 | −0.149 | −0.121 |
| (3) Flanker Δ RT congruency | 0.033 | −0.025 | 0.062 | −0.081 | 0.050 | −0.045 | −0.118 |
| (4) WCST perseverative errors | −0.137 | 0.164 | −0.138 | 0.036 | −0.050 | 0.037 | 0.157 |
| (5) WCST non-perseverative errors | 0.030 | 0.349* | −0.177 | 0.090 | −0.065 | −0.181 | 0.139 |
| (6) WCST set failure | 0.089 | −0.075 | −0.041 | −0.040 | 0.073 | −0.042 | −0.264* |
| (7) WCST total errors | −0.060 | 0.296* | −0.194[†] | 0.069 | −0.098 | −0.104 | 0.170 |
| (8) BDS final score | 0.130 | 0.141 | −0.248* | −0.006 | −0.058 | −0.168 | −0.109 |
| (9) BDS proportion correct | 0.021 | 0.147 | −0.125 | −0.118 | 0.005 | −0.118 | −0.047 |

All home measures and subjective noise measures are ordinal, other than the number of inhabitants. Total income has a scale of 1 to 9, while all other ordinal measures have a scale of 1 to 4. $N = 82–125$.

*Significant. [†]Near significant.

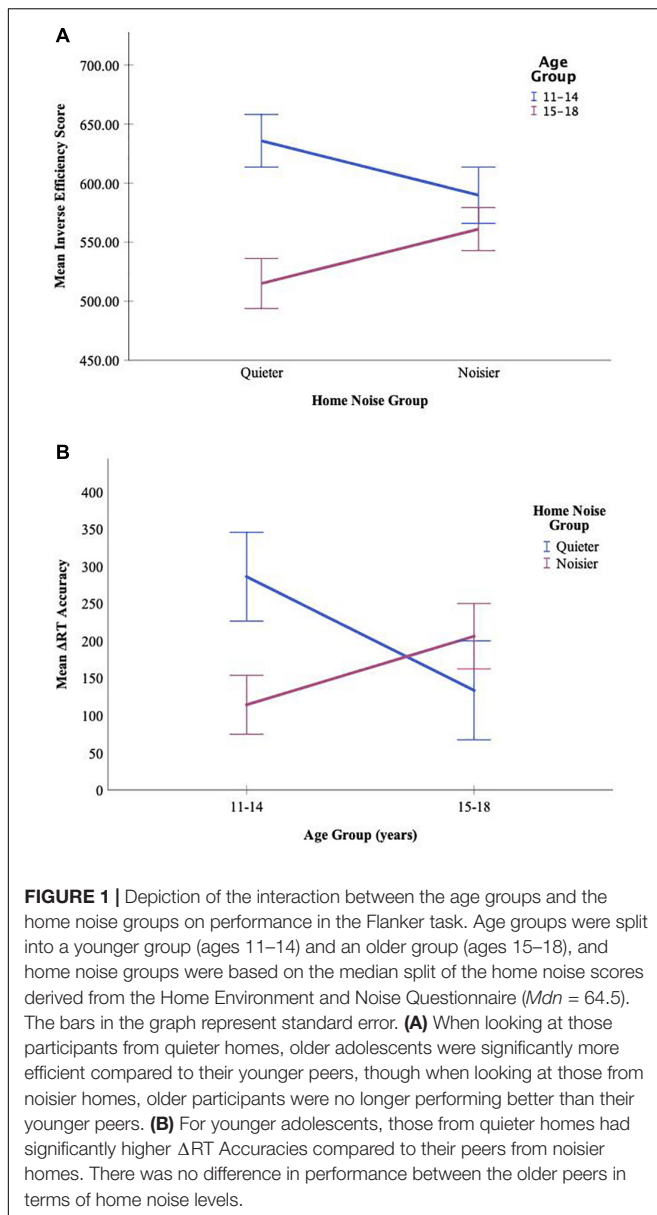
subjective noise characteristics and task performance, only a higher reported annoyance to noise correlated with less WCST failures to maintain set.

Experimental Noise and Executive Function Task Performance

According to plan, participant task scores and the noise questionnaire scores that were above or below three standard deviations from the mean were removed before analysis (fewer than four data points were removed for each variable). See **Table 5** for the included ages and genders across experimental conditions and home noise groupings. Following the same plan, order effects from background noise habituation were also checked, since participants heard the same noise soundtrack during each task. Task order and noise condition were run through a MANOVA which included all task scores. There were no main effects of noise condition nor any significant interactions between task order and noise condition. The only main effect of task order was for the Flanker IES, $F(2,75) = 5.45$, $p = 0.006$, $\eta^2_p = 0.125$. A Tukey *post hoc* test revealed that those who completed the Flanker task as their second task had a significantly higher IES score ($M = 629.36$, $SD = 110.83$) than those who completed it as their first ($M = 554.54$, $SD = 100.89$, $p = 0.012$) or third

($M = 551.01$, $SD = 79.56$, $p = 0.018$) task. There was no significant difference in IES between those who completed the Flanker task first or last. Thus, regardless of being in the White Noise or the Environmental Noise condition, those who completed the Flanker task second seemed to be less efficient at completing the task than those who completed the Flanker task first or last.

With no significant habituation to the noise found, as pre-registered, the effects of noise condition, home noise scores and age on task performance were looked at. A 2 (noise condition: white, environmental) \times 2 (home noise: quieter, noisier) \times 2 (age: 11–14, 15–18) MANCOVA was run looking at all task scores, with all predictors being between-subject variables. A multivariate analysis was run instead of the planned separate univariate tests for each task score to enable equivalent sample sizes and participants be included in each analysis, ensuring comparability between the findings. Although country of residence (United Kingdom or United States) did not significantly correlate with the home measures, subjective noise measures, and task performance, many cultural differences could still be present that were not accounted for that could interact with the effect of noise on EF task performance. Thus, country of residence was added into the analysis as a covariate. Changes to the SPSS syntax were made (see **Supplementary Appendix B**)



in order to use the covariate influenced adjusted means in the *post hoc* analyses, and all pairwise comparisons were Bonferroni corrected. No significant effect of country was found.

Flanker

The Flanker task consisted of IES, ΔRT Accuracy, and ΔRT Congruency scores. Results revealed no main effects of noise condition or home noise, but there was a significant main effect of age for the IES score, with older participants ($M = 533.20$, $SD = 88.21$) having a more efficient score than the younger participants ($M = 612.52$, $SD = 107.91$), $F(1,73) = 12.02$, $p = 0.001$, $\eta^2_p = 0.141$. There were no significant interactions between noise condition and home noise or age; however, the IES score had a significant interaction between home noise and age, $F(1,73) = 5.50$, $p = 0.022$, $\eta^2_p = 0.070$. Further analyses revealed that if from a quieter home, the older adolescents performed

more efficiently ($M = 509.71$, $SD = 84.76$) than the younger adolescents ($M = 641.08$, $SD = 106.70$), $p < 0.001$, but this advantage was no longer present if they were from a noisier home ($M_{15-18} = 556.69$, $SD_{15-18} = 87.36$ and $M_{11-14} = 583.96$, $SD_{11-14} = 106.61$), $t(41) = 0.97$, $p = 0.34$ (see **Figure 1A**).

Furthermore, the ΔRT Accuracy score had the same significant interaction between home noise and age, $F(1,73) = 3.97$, $p = 0.050$, $\eta^2_p = 0.052$. Further analyses revealed a different direction, however, whereby in the younger age group, adolescents from quieter homes had a higher ΔRT Accuracy ($M = 285.72$, $SD = 285.42$) compared to their peers from noisier homes ($M = 132.85$, $SD = 176.67$), $p = 0.045$. There was no difference though in ΔRT Accuracy depending on home noise in the older adolescents ($M_{quieter} = 146.33$, $SD_{quieter} = 265.16$ and $M_{noisier} = 209.79$, $SD_{noisier} = 210.28$), $p = 0.42$ (see **Figure 1B**). Lastly, there were no significant three-way interactions between noise condition, home noise, and age for the Flanker task.

Wisconsin Card Sorting Test

The WCST was scored based on the number of perseverative errors, non-perseverative errors, failure to maintain set, and total errors. The analyses revealed no main effects of noise condition, home noise, or age. While there were no significant interactions between home noise and noise condition or age, there were two near significant interactions between noise condition and age: Perseverative errors [$F(1,73) = 3.23$, $p = 0.076$, $\eta^2_p = 0.042$] and total errors [$F(1,73) = 3.15$, $p = 0.080$, $\eta^2_p = 0.041$]. The younger adolescent group showed a trend toward more perseverative errors in the Environmental Noise condition ($M = 5.52$, $SD = 4.72$) compared to those in the White Noise condition ($M = 3.08$, $SD = 2.25$), $p = 0.064$; however, there was no difference between the noise conditions for the older adolescents ($M_{enviro} = 4.26$, $SD_{enviro} = 3.99$ and $M_{white} = 5.19$, $SD_{white} = 4.73$), $p = 0.49$. Similarly, the younger adolescents also trended toward making more total errors in the Environmental Noise condition ($M = 20.12$, $SD = 11.55$) than in the White Noise condition ($M = 15.09$, $SD = 5.32$), $p = 0.088$. Again, there was no difference in total errors made between the two noise conditions for the older adolescents ($M_{enviro} = 15.97$, $SD_{enviro} = 6.68$ and $M_{white} = 18.38$, $SD_{white} = 10.38$), $p = 0.43$. Finally, no three-way interactions were found.

Backward Digit Span

The BDS task was evaluated based on final score and proportion correct. Analyses revealed no significant main effects or interactions.

DISCUSSION

The main purpose of the current study was to run an independent online experiment looking at the effect of the home, and in particular in-home noise, on adolescent EF. The EF tasks used reflect the skills that are frequently needed within academic learning; therefore, any effects on their EF task performance could indicate the potential impact that noise may be having on their in-home learning during the pandemic. Another more exploratory avenue of the current study was to understand

TABLE 5 | Breakdown of subject characteristics within both experimental Noise Condition and Home Noise grouping based on gender and age.

| | Noise Condition | | | | Home Noise | | | | Total Sample | |
|--------------------|-----------------|-------|---------------|-------|------------|-------|---------|-------|--------------|-------|
| | White | | Environmental | | Quieter | | Noisier | | N | M |
| | N | M | N | M | N | M | N | M | | |
| Gender | 38 | | 42 | | 38 | | 42 | | 80 | |
| Female | 18 | | 26 | | 18 | | 26 | | 44 | |
| Male | 20 | | 16 | | 20 | | 16 | | 36 | |
| Age (years) | 39 | 14.64 | 43 | 14.76 | 39 | 14.34 | 43 | 15.03 | 82 | 14.70 |
| 11 | 7 | 11.55 | 4 | 11.48 | 9 | 11.50 | 1 | 11.62 | 11 | 11.52 |
| 12 | 3 | 12.33 | 6 | 12.46 | 5 | 12.25 | 4 | 12.62 | 9 | 12.42 |
| 13 | 6 | 13.42 | 4 | 13.58 | 4 | 13.33 | 6 | 13.58 | 10 | 13.48 |
| 14 | 3 | 14.17 | 10 | 14.35 | 5 | 14.42 | 8 | 14.24 | 13 | 14.31 |
| 15 | 7 | 15.27 | 4 | 15.42 | 4 | 15.54 | 7 | 15.20 | 11 | 15.33 |
| 16 | 7 | 16.22 | 6 | 16.18 | 4 | 16.27 | 9 | 16.17 | 13 | 16.20 |
| 17 | 1 | 17.67 | 8 | 17.43 | 5 | 17.47 | 4 | 17.44 | 9 | 17.46 |
| 18 | 5 | 18.38 | 1 | 18.08 | 3 | 18.19 | 3 | 18.47 | 6 | 18.33 |
| Age groups | | | | | | | | | | |
| Younger (11–14) | 19 | 12.68 | 24 | 13.27 | 23 | 12.62 | 20 | 13.46 | 43 | 13.01 |
| Older (15–18) | 20 | 16.50 | 19 | 16.65 | 16 | 16.82 | 23 | 16.40 | 39 | 16.57 |

how factors determining the home environment relate to subjective perceptions of noise, and how these both might relate to adolescent EF.

Correlations Between the Home Environment, Perception of Noise, and Executive Function

In terms of the home environment, it is clear that the pandemic led to population shifts. The recorded decreases and increases of inhabitants in the home could represent both the more vulnerable inhabitants moving out of the home to be more protected on their own, as well as separate households grouping together to support each other throughout the pandemic and ongoing lockdowns. While the specific reasons for shifting homes were not directly recorded, overall, there was a small but significant increase in the number of inhabitants living in the home during the pandemic compared to before, indicating that the core make-up of a home was affected by the pandemic.

To get a better understanding of the adolescents' home environments, the current study further measured total family income, in-home noise levels, number of inhabitants, and dwelling type. The home noise score that was derived from the questionnaire positively correlated with dwelling type, indicating that the more collective the dwelling, the higher their in-home noise scores. While we did not directly measure noise levels in the participants' homes, this finding does follow the same conclusion as Pujol et al. (2012) who directly measured the in-home noise levels of a similar demographic (20% detached dwellings and 80% collective dwellings) over 8 days. It was further found that more inhabitants in the home also coincided with living in less collective dwellings, consistent with larger families needing a bigger home. What is interesting, though, is that unlike Pujol et al. (2012), a correlation between in-home noise levels and

number of inhabitants was not found. While we cannot conclude from the current results that more people dwelling together during the pandemic increased the noise levels, having a more direct measure of home noise and any changes in the noise levels from before the pandemic to during the pandemic, along with the shifts in household numbers, might better capture this relationship. Another divergence from previous literature (Dale et al., 2015; Casey et al., 2017) was found, where a lower income did not correlate with higher in-home noise levels. Perhaps a more concise depiction of SES, as was originally planned in the study, would have been better able to measure this.

Subjective perceptions of noise were also recorded, including adolescents' general annoyance to noise compared to their peers, their preference for background noise when studying, their perception of the noise level in the room they study in, and their annoyance with the in-room noise. The two significant correlations between these were that the higher they reported their annoyance with in-room noise, the noisier they reported their room to be and the more annoyed to noise in general they reported being compared to their peers. This coincides with the literature, where higher noise levels correlated with higher rates of annoyance (Lundquist et al., 2000; Ali, 2013; Minichilli et al., 2018). In terms of how the home measures correlated with subjective noise measures, unsurprisingly, higher home noise scores correlated with a perception of higher in-room noise levels. Furthermore, those with higher in-home noise scores were more likely to report a preference for a noisier background environment when studying, and also more annoyance with in-room noise. Evidence for the possible explanation that those from noisier homes either develop a preference for noise *or* become more annoyed by noise, was not found. Thus, the findings suggest that those from noisier homes both prefer to have more noise in the background when studying yet are also more annoyed

by in-room noise. Perhaps, those from noisier homes find it more difficult to work in silence and require some noise in the background to match the environment that they are most used to, but that these same noisier homes are more likely to have particular sound sources that are more annoying than would be found in a quieter home. This would align with Connolly et al.'s (2013) finding where adolescents reported different levels of annoyance depending on the type of sound present, meaning that further research into the varying effects of specific noise sources within the home is needed. Of note, as the home noise score was based on the reporting of the frequency of specific in-home sounds, this score is therefore susceptible to subjective perceptions of noise and thus it is not surprising that the noise score correlated with subjective noise measures.

Contrary to the literature showing an effect of annoyance by age within school environments (Ali, 2013; Connolly et al., 2013; Minichilli et al., 2018), the current study did not find a difference in home noise-induced annoyance in younger versus older adolescents. However, the cited studies took place within school settings, and not within a home-learning environment during a pandemic. There are many possible reasons for which annoyance levels might now differ, from familiarity with home noises to frustration with trying to learn in novel circumstances. Task type and cognitive demand at the time of reporting have also been suggested to mediate the effect of age on noise annoyance (Connolly et al., 2013). Lastly, as age did not correlate with the home measures, we can conclude that while our age range was large, participants within each age-point included came from diverse homes, strengthening the generalization of our findings. With the current evidence that reports of noise-induced annoyance relate to dwelling type and do not relate to age, it is clear that the findings on annoyance from the school literature cannot fully capture what is happening in the home and how adolescents are being impacted by in-home noise during the pandemic.

We further looked at how measures of the home environment and subjective measures of noise might relate to adolescent performance on the three EF tasks. While total family income did not correlate with task performance, number of inhabitants in the home during the pandemic related to both the Flanker and the WCST tasks. Adolescents in a home with more inhabitants trended toward having a higher Flanker Δ RT Accuracy score and had significantly more non-perseverative and total errors on the WCST. Furthermore, those adolescents who live in a more collective dwelling trended toward being more efficient on the Flanker task and having significantly lower Δ RT Accuracies. They also trended toward being more likely to have fewer overall errors on the WCST and were significantly more likely to have a worse BDS final score. While it appears that overall a higher number of inhabitants correlates with worse performance on the WCST- a task that involves shifting, working memory, and inhibition- and a potential difference in response time behavior on the Flanker inhibitory task, the relationship with dwelling type is not as clear.

It seems that while there is evidence that being from a more collective dwelling positively correlates with better task performance on the Flanker and WCST, this also negatively

correlates with performance on the BDS task. However, as will be discussed later, data from the BDS task may not be reliable and thus might explain this conflicting finding. This might then infer that overall, coming from a home with closer and more neighbors may be linked to better adolescent EF. Lastly, in terms of subjective measures of noise and EF task performance, being more annoyed by in-room noise correlating with less set failures on the WCST was the only significant result. While the direct relationship between EF task performance and both measures of the home environment and subjective noise cannot be inferred, it is clear that factors strongly determining the home environment are linked to adolescent EF abilities; the link between individual differences in the subjective experience with noise and EF is less evident.

Effect of Noise on Executive Function

It was hypothesized that there would be a direct effect of the audio recording condition on task performance. While we did not find this overarching effect, when splitting participants into younger and older adolescent age groups, results showed a clear *trend* on the WCST whereby the younger adolescents were making more perseverative and total errors in the presence of the environmental noise, compared to those simply in the white noise condition. Connolly et al. (2019) did find a significant interaction of age and school environmental background noise, though the study specifically looked at reading ability and found varying age effects at different noise levels. While the current results just missed statistical significance, the evident and identical direction of the trends mean that while a strong conclusion cannot currently be made, nor can these results be discounted. Future research should look at how changing and dynamic sounds often found in noisier homes directly impact on learning.

A main effect of age was found when looking at the Flanker task efficiency score, which takes into account speed of reaction time and accuracy, with older adolescents performing more efficiently than their younger peers. Furthermore, while there was not a clear prediction for the effect of the in-home noise on task performance, when splitting the participants into their separate age groups, we did find significant results for this same Flanker efficiency score. When looking at those from quieter homes, older adolescents still demonstrated more efficiency on the Flanker task than their younger peers, but this advantage disappeared when looking at those from noisier homes. The overall finding that older adolescents perform more efficiently on this EF task regardless of noise follows previous research (for a review, see Ridderinkhof et al., 2021). What is, however, unexpected and remarkable, is the finding that when taking into account individual differences, such as the noise levels that the adolescent experiences on a daily basis at home, the older adolescents no longer show this developmental advantage in their performance on the task.

Another interaction between the effects of in-home noise and age on the Flanker task was found for the Δ RT Accuracy score. When looking exclusively at the younger adolescents, those who came from noisier homes had higher Δ RT Accuracy scores compared to their peers from noisier homes. This implies that if they experience more in-home noise, they are more likely

to have similar reaction times for both correct and incorrect trials, whereas those from quieter homes clearly have a behavioral difference in their response times depending on accuracy. With no differences found for the older adolescents, it is apparent that only the younger adolescents are impacted by the long exposure to noise in this instance.

Overall, we can infer from these findings that regardless of the noise recording being played during the experiment, the noise that adolescents are frequently surrounded by in their home is having long-term effects on their EF. This finding, therefore, extends past pandemic-specific circumstances as it implies that regardless of the environment that they are learning in, be it their home or their school, coming from a home with higher noise levels can have disadvantageous effects for both older and younger adolescents.

Of note, it was predicted that there would be an interaction between noise condition and home noise, where those who experience higher in-home noise on a daily basis would do better in the environmental noise condition than those from quieter homes; however, no evidence was found for this on any of the tasks. Therefore, it does not seem that experience with in-home noise translates to a novel learning situation with similar noise. One potential limitation of the study that could explain why there was not an effect of audio recording condition on task performance could be that since participants heard the same audio recording repeated for each task, over time, they could have habituated to the noise and thus their performance would no longer have been affected by it. However, no order effects based on noise condition were found, meaning that the participants did not become habituated to the noise. Another possible explanation is that while the environmental audio was created to depict a naturalistic noisy home, homes can vary on specific sound sources and frequency of sounds; thus, perhaps these intricacies that make up their in-home noise experience need to be matched in the audio in order for them to perform better compared to their peers from quieter homes. For example, while some participants from a noisy home may frequently hear planes overhead, peers from equivalently noisy homes may never hear planes, and thus would get more distracted by this sound source while completing the tasks. Thus, as mentioned previously, further research into the varying impact of specific sound sources within the home is needed.

Going back to the results on order effects, an overall effect of task order regardless of background noise was uncovered, with a higher IES score when the Flanker task was completed second. This potentially could be explained by research showing “inhibitory fatigue,” whereby when completing two consecutive inhibition tasks, performance on the second is likely to be poorer than if a different task had preceded it (Diamond, 2013). However, because the WCST preceded the Flanker task both when the Flanker was completed second and third, the finding of decreased efficiency when completed second cannot be due to this. As the order effect found did not interact with background noise, age, and home noise when these were checked, while there is no clear explanation for the finding, it was concluded that it had no influence on the current findings.

Limitations and Future Directions

As the current study was designed and completed during a pandemic, it is important to highlight the limitations that were present in the current design. Importantly, while the two noise conditions used offered the ability to understand the influence of environmental noise, a true control condition without any noise would have been preferable. For instance, Helps et al. (2014) covaried for performance in a no-noise condition to determine the true effect of different levels of white noise on performance when testing children in a school room setting. While this may be feasible for certain designs where the children are all tested in the same environment and are exposed to the same environmental noise in the room, this was not feasible to implement in the current study. It is important to note as well that white noise has been found to influence children’s EF task performance, with certain levels of white background noise aiding low-attentive children and hindering high attentive children (Söderlund et al., 2010; Helps et al., 2014). Future research looking at the differences between environmental noise, white noise, and no noise would help to better understand and interpret the current findings. Additionally, the noise questionnaire used here has not been validated against true measures of in-home noise levels, and as previously mentioned, it is susceptible to subjective perceptions of noise. Without a direct measure of noise, the current study was not able to disentangle objective and subjective effects of noise, though with learning being such a multifaceted construct, it is likely that both play an important role. In terms of the participants, while neurotypical adolescents were advertised for during recruitment, further checks should be implemented in future to ensure that other factors linked to EF ability and noise sensitivity, such as autism (Kouklari et al., 2018; Schwartz et al., 2020), are not influencing the results.

Of further note, as the EF tasks used were a proxy for the cognitive demands often found within academic learning, the current study is a first step toward understanding the direct effect of in-home noise on home-learning. Further research is very much needed to fully understand the extent to which the pandemic has affected students within secondary education. For instance, a recent study by Muzi et al. (2021) looking at adolescent wellbeing during the pandemic found an increase in problematic social media usage, which was then further linked to higher rates of attentional and other emotional-behavioral problems. Future work should therefore look at the interplay between noise and social media distractions and its effects on adolescent attention, EF, and learning, especially within the context of the pandemic. The authors further highlight how adolescents with insecure attachment may be more susceptible to the fear and isolation brought about by the pandemic (Muzi et al., 2021). With attachment being linked to both EF (Escobar et al., 2013) and the home environment (Klemfuss et al., 2018), it would be important to take into account how attachment may be moderating the relationship between in-home noise and EF task performance, particular when considering that certain social-induced noises (e.g., a parent scolding a sibling) may have a different effect and may be more linked to attachment than a non-social noise (e.g., the washing machine running).

Advantages and Disadvantages of Independently Run Online Research

The potential advantages of independently run online research is vast, both for researchers and for the inclusion of heterogeneous participants. Importantly, for researchers, independent online research can increase productivity by reducing the many months, and sometimes years that are spent collecting data. In addition, it allows for research groups with less funding for bringing participants into the lab, or indeed smaller spaces, to conduct large-scale projects. Furthermore, projects and ideas are sometimes limited due to the time imposed by data collection, and an increase in the online tools available to conduct high caliber research can significantly change this. Crucially though, there are certain tasks and forms of research that will not be able to be translated to an online and/or independent format. The BDS task used in the current study is a prime example. While it is a popular and well validated working memory task within the field due to its use in the standardized Wechsler Intelligence Scale for Children (Wechsler, 2014), it does not translate well to an online and independent format. Regardless of telling participants to not write down the numbers, it is likely that many participants still did this, potentially explaining the current lack of findings for this task. Thus, as the BDS does not seem to be adaptable, conclusions for BDS performance have not been made in the current study. With no easy way of controlling for this limitation, the future use of this task in independently run online environments is not advised.

In terms of online research helping with participant heterogeneity, as recruitment is not limited to a specific location, it has the potential to recruit a much more diverse participant pool. Location based research tends to only attract families of higher socioeconomic status that have the time and financial freedom to travel and spend a few hours at the lab, making it difficult to break the W.E.I.R.D cycle of data collection. Running independent online research can also enable more global research, as time zones are no longer a constraint. Of note though, simply translating research to an online format does not automatically lead to a more heterogeneous sample, as can be seen in the current sample, and careful steps still need to be taken to include a more diverse sample. Furthermore, issues with collecting sociodemographic information arose. In the current study, it was evident that these independent adolescents occasionally completed the sociodemographic questionnaire with their own information rather than their parents', reducing the data that we could interpret. We did find that including the option to select "Do not know," as we had for total family income, reduced the reporting of incorrect data. Thus, by making it abundantly clear who the question is referring to, as well as giving participants an option to opt out in case their parent is not accessible at the time of completing the questionnaire, will ensure accurate sociodemographic data collection.

Naturally, with an independently run online study, there is less researcher control over the testing environment. Steps, however, can be taken to ensure experimental rigor. For instance, as auditory stimuli were key for the current experiment, an objective

headphone screening task worked well to guarantee good hearing ability, working headphones, and that the participants were wearing the headphones. This did, however, mean that before data processing, 19 participants were already excluded. Furthermore, additional pre-processing steps were included in the current study to help ensure high data quality. Participants or individual data points were excluded based on loading delays, time taken to complete the task, response time, not following the rules, audio consistently restarting, audio stopping before the end of the task, and self-reported audio issues. Unfortunately, this inevitably means greater data loss, with many of these exclusions not being necessary or as common during in-lab testing. Fortunately, with online data collection being faster, including more participants is easy enough to ensure high data quality. So while some control of the testing environment is lost in online and independently run studies, steps can be implemented to resolve these issues and allow researchers to reap the many benefits that this methodology enables.

CONCLUSION

Overall, the current study clearly demonstrates that the home environment influences the subjective perception of noise. In particular, we found converging evidence with the school literature that higher levels of noise correlate with higher rates of annoyance in adolescents. Furthermore, while we did not find a significant direct effect of background noise on EF task performance, actual in-home noise levels significantly affected task performance. Regardless of the background audio presented while completing the task, both younger and older adolescents showed evidence that consistently being in a noisy home impacted their EF task performance. With in-home noise levels having long-term effects on EF, it is clear that more research needs to be done to better understand the influence that the home environment may be having on learning within the home, as well as within schools.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the School of Sciences Ethics Committee at Birkbeck, University of London (reference number: 192071). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin, or by the participant themselves if they were 16 years of age or older.

AUTHOR CONTRIBUTIONS

BC and NK contributed to the conception and design of the study. BC created the study, recruited participants, pre-processed

the data, performed the statistical analysis, and wrote the first draft of the manuscript. NK wrote sections of the manuscript. Both authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This research was funded by an Economic and Social Research Council UBEL studentship (grant reference: ES/P000592/1).

REFERENCES

- Ali, S. A. A. (2013). Study effects of school noise on learning achievement and annoyance in Assiut city. *Egypt. Appl. Acoust.* 74, 602–606. doi: 10.1016/j.apacoust.2012.10.011
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: an online behavioral experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x
- Best, J. R., Miller, P. H., and Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learn. Individ. Differ.* 21, 327–336. doi: 10.1016/j.lindif.2011.01.007
- Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., et al. (2018). What accounts for individual differences in susceptibility to the McGurk effect? *PLoS One* 13:e0207160. doi: 10.1371/journal.pone.0207160
- Cameron-Blake, E., Tatlow, H., Wood, A., Hale, T., Kira, B., Petherick, A., et al. (2020). *Variation in the Response to COVID-19 Across the Four Nations of the United Kingdom. Blavatnik School of Government Working Paper*. Available online at: www.bsg.ox.ac.uk/covidtracker
- Casey, J. A., Morello-Frosch, R., Mennitt, D. J., Frstrup, K., Ogburn, E. L., and James, P. (2017). Race/ethnicity, socioeconomic status, residential segregation, and spatial variation in noise exposure in the contiguous United States. *Environ. Health Perspect.* 125:77017. doi: 10.1289/EHP898
- Cianchetti, C., Corona, S., Foscoliano, M., Contu, D., and Sannio-Fancello, G. (2007). Modified Wisconsin Card sorting Test (MCST, MWCSST): normative data in children 4–13 Years old, according to classical and new types of scoring. *Clin. Neuropsychol.* 21, 456–478. doi: 10.1080/13854040600629766
- Connolly, D., Dockrell, J., Shield, B., Conetta, R., Mydlarz, C., and Cox, T. (2019). The effects of classroom noise on the reading comprehension of adolescents. *J. Acoust. Soc. Am.* 145, 372–381. doi: 10.1121/1.5087126
- Connolly, D. M., Dockrell, J. E., Shield, B. M., Conetta, R., and Cox, T. J. (2013). Adolescents' perceptions of their school's acoustic environment: the development of an evidence based questionnaire. *Noise Health* 15:269. doi: 10.4103/1463-1741.113525
- Dale, L. M., Goudreau, S., Perron, S., Ragetti, M. S., Hatzopoulou, M., and Smargiassi, A. (2015). Socioeconomic status and environmental noise exposure in Montreal, Canada. *BMC Public Health* 15:205. doi: 10.1186/s12889-015-1571-2
- Department for Education and Skills (2015). *Building Bulletin 93 Acoustic Design of Schools: Performance Standards*. London: The Stationary Office.
- Diamond, A. (2013). Executive functions. *Ann. Rev. Psychol.* 64, 135–168. doi: 10.1146/annurev-psych-113011-143750
- Dumas, J. E., Nissley, J., Nordstrom, A., Smith, E. P., Prinz, R. J., and Levine, D. W. (2005). Home chaos: sociodemographic, parenting, interactional, and child correlates. *J. Clin. Child Adolesc. Psychol.* 34, 93–104. doi: 10.1207/s15374424jccp3401_9
- Enmarker, I., and Boman, E. (2004). Noise annoyance responses of middle school pupils and teachers. *J. Environ. Psychol.* 24, 527–536. doi: 10.1016/j.jenvp.2004.09.005
- Escobar, M. J., Rivera-Rei, A., Decety, J., Huepe, D., Cardona, J. F., Sigman, M., et al. (2013). Attachment patterns trigger differential neural signature of emotional processing in adolescents. *PLoS One* 8:e70247. doi: 10.1371/journal.pone.0070247
- Fink, D. J. (2017). What is a safe noise level for the public? *Am. J. Public Health* 107, 44–45. doi: 10.2105/AJPH.2016.303527
- Guski, R., Schreckenberger, D., and Schuemer, R. (2017). WHO environmental noise guidelines for the European region: a systematic review on environmental noise and annoyance. *Int. J. Environ. Res. Public Health* 14:1539. doi: 10.3390/ijerph14121539
- Hale, T., Atav, T., Hallas, L., Kira, B., Phillips, T., Petherick, A., et al. (2020). *Variation in US States' Responses to COVID-19. Blavatnik School of Government Working Paper*. Available online at: www.bsg.ox.ac.uk/covidtracker
- Helps, S. K., Bamford, S., Sonuga-Barke, E. J., and Söderlund, G. B. (2014). Different effects of adding white noise on cognitive performance of sub-, normal and super-attentive school children. *PLoS One* 9:e112768. doi: 10.1371/journal.pone.0112768
- Huizinga, M., and van der Molen, M. W. (2007). Age-group differences in set-switching and set-maintenance on the Wisconsin card sorting task. *Dev. Neuropsychol.* 31, 193–215. doi: 10.1080/87565640701190817
- Imburgio, M. J., Banica, I., Hill, K. E., Weinberg, A., Foti, D., and MacNamara, A. (2020). Establishing norms for error-related brain activity during the arrow Flanker task among young adults. *NeuroImage* 213:116694. doi: 10.1016/j.neuroimage.2020.116694
- Jacob, R., and Parkinson, J. (2015). The potential for school-based interventions that target executive function to improve academic achievement: a review. *Rev. Educ. Res.* 85, 512–552. doi: 10.3102/0034654314561338
- King, L. S., Dennis, E. L., Humphreys, K. L., Thompson, P. M., and Gotlib, I. H. (2020). Cross-sectional and longitudinal associations of family income-to-needs ratio with cortical and subcortical brain volume in adolescent boys and girls. *Dev. Cogn. Neurosci.* 44:100796. doi: 10.1016/j.dcn.2020.100796
- Klatte, M., Bergström, K., and Lachmann, T. (2013). Does noise affect learning? A short review on noise effects on cognitive performance in children. *Front. Psychol.* 4:578. doi: 10.3389/fpsyg.2013.00578
- Klemfuss, J. Z., Wallin, A. R., and Quas, J. A. (2018). Attachment, household chaos, and children's health. *Fam. Syst. Health* 36:303. doi: 10.1037/fsh0000303
- Kouklari, E. C., Tsermentseli, S., and Monks, C. P. (2018). Hot and cool executive function in children and adolescents with autism spectrum disorder: cross-sectional developmental trajectories. *Child Neuropsychol.* 24, 1088–1114. doi: 10.1080/09297049.2017.1391190
- Lipsey, M. W., Nesbitt, K. T., Farran, D. C., Dong, N., Fuhs, M. W., and Wilson, S. J. (2017). Learning-related cognitive self-regulation measures for prekindergarten children: a comparative evaluation of the educational relevance of selected measures. *J. Educ. Psychol.* 109, 1084–1102. doi: 10.1037/edu0000203
- Lundquist, P., Holmberg, K., and Landstrom, U. (2000). Annoyance and effects on work from environmental noise at school. *Noise Health* 2:39. doi: 10.1121/2.0000641
- MacArthur Research Network on SES and Health (2008). *Sociodemographic Questionnaire*. Available online at: <https://macses.ucsf.edu/research/socialenviron/sociodemographic.php> (accessed May 22, 2020).
- Massonnié, J. (2020). *Understanding the Impact of Classroom Noise on Children's Learning and Well-Being, and its Modulation by Executive Functions*. Doctoral dissertation. London: Birkbeck, University of London.

ACKNOWLEDGMENTS

Thank you to the administration team at the Centre for Brain and Cognitive Development and to the support team at Gorilla. Thank you to all families who took part in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.715301/full#supplementary-material>

- Massonnié, J., Frassetto, P., Mareschal, D., and Kirkham, N. Z. (2020). Learning in noisy classrooms: children's reports of annoyance and distraction from noise are associated with individual differences in mind-wandering and switching skills. *Environ. Behav.* doi: 10.1177/0013916520950277 [Epub ahead of print].
- McClelland, M. M., and Cameron, C. E. (2019). Developing together: the role of executive function and motor skills in children's early academic lives. *Early Child. Res. Q.* 46, 142–151. doi: 10.1016/j.jecresq.2018.03.014
- Minichilli, F., Gorini, F., Ascari, E., Bianchi, F., Coi, A., Fredianelli, L., et al. (2018). Annoyance judgment and measurements of environmental noise: a focus on Italian secondary schools. *Int. J. Environ. Res. Public Health* 15:208. doi: 10.3390/ijerph15020208
- Mullane, J. C., Corkum, P. V., Klein, R. M., and McLaughlin, E. (2009). Interference control in children with and without ADHD: a systematic review of Flanker and Simon task performance. *Child Neuropsychol.* 15, 321–342. doi: 10.1080/09297040802348028
- Muzi, S., Sansò, A., and Pace, C. S. (2021). What's happened to Italian adolescents during the COVID-19 pandemic? A preliminary study on symptoms, problematic social media usage, and attachment: relationships and differences with pre-pandemic peers. *Front. Psychiatry* 12:543. doi: 10.3389/fpsy.2021.590543
- Parmentier, F. B. (2014). The cognitive determinants of behavioral distraction by deviant auditory stimuli: a review. *Psychol. Res.* 78, 321–338. doi: 10.1007/s00426-013-0534-4
- Passchier-Vermeer, W., and Passchier, W. F. (2000). Noise exposure and public health. *Environ. Health Perspect.* 108(suppl 1), 123–131. doi: 10.1289/ehp.00108s1123
- Piper, B. J., Li, V., Eiwaz, M. A., Kobel, Y. V., Benice, T. S., Chu, A. M., et al. (2012). Executive function on the psychology experiment building language tests. *Behav. Res. Methods* 44, 110–123. doi: 10.3758/s13428-011-0096-6
- Pujol, S., Berthillier, M., Defrance, J., Lardies, J., Levain, J.-P., Petit, R., et al. (2014). Indoor noise exposure at home: a field study in the family of urban schoolchildren. *Indoor Air* 24, 511–520. doi: 10.1111/ina.12094
- Pujol, S., Berthillier, M., Defrance, J., Lardiès, J., Petit, R., Houot, H., et al. (2012). Urban ambient outdoor and indoor noise exposure at home: a population-based study on schoolchildren. *Appl. Acoust.* 73, 741–750. doi: 10.1016/j.apacoust.2012.02.007
- Ridderinkhof, K. R., Wylie, S. A., van den Wildenberg, W. P. M., Bashore, T. R., and van der Molen, M. W. (2021). The arrow of time: advancing insights into action control from the arrow version of the Eriksen flanker task. *Atten. Percept. Psychophys.* 83, 700–721. doi: 10.3758/s13414-020-02167-z
- Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., et al. (2004). Development of attentional networks in childhood. *Neuropsychologia* 42, 1029–1040. doi: 10.1016/j.neuropsychologia.2003.12.012
- Schwartz, S., Wang, L., Shinn-Cunningham, B. G., and Tager-Flusberg, H. (2020). Atypical perception of sounds in minimally and low verbal children and adolescents with autism as revealed by behavioral and neural measures. *Autism Res.* 13, 1718–1729. doi: 10.1002/aur.2363
- Shield, B., Conetta, R., Dockrell, J., Connolly, D., Cox, T., and Mydlarz, C. (2015). A survey of acoustic conditions and noise levels in secondary school classrooms in England. *J. Acoust. Soc. Am.* 137, 177–188. doi: 10.1121/1.4904528
- Shield, B. M., and Dockrell, J. E. (2003). The effects of noise on children at school: a review. *Build. Acoust.* 10, 97–116. doi: 10.1260/135101003768965960
- Söderlund, G. B., Sikström, S., Loftesnes, J. M., and Sonuga-Barke, E. J. (2010). The effects of background white noise on memory performance in inattentive school children. *Behav. Brain Funct.* 6:55. doi: 10.1186/1744-9081-6-55
- Stevens, C., and Bavelier, D. (2012). The role of selective attention on academic foundations: a cognitive neuroscience perspective. *Dev. Cogn. Neurosci.* 2, S30–S48. doi: 10.1016/j.dcn.2011.11.001
- van Leeuwen, M., Hoekstra, R. A., and Boomsma, D. I. (2007). Endophenotypes for intelligence in children and adolescents. *Intelligence* 35, 369–380. doi: 10.1016/j.intell.2006.09.008
- Wechsler, D. (2014). *WISC-V: Technical and Interpretive Manual*. Bloomington, MN: NCS Pearson, Incorporated.
- Wells, E. L., Kofler, M. J., Soto, E. F., Schaefer, H. S., and Sarver, D. E. (2018). Assessing working memory in children with ADHD: minor administration and scoring changes may improve digit span backward's construct validity. *Res. Dev. Disabil.* 72, 166–178. doi: 10.1016/j.ridd.2017.10.024

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Chere and Kirkham. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Framework for Online Experimenter-Moderated Looking-Time Studies Assessing Infants' Linguistic Knowledge

Desia Bacon[†], Haley Weaver^{†*} and Jenny Saffran

Infant Learning Lab, Department of Psychology, Waisman Center, University of Wisconsin-Madison, Madison, WI, United States

OPEN ACCESS

Edited by:

Natasha Kirkham,
Birkbeck, University of London,
United Kingdom

Reviewed by:

LouAnn Gerken,
University of Arizona,
United States
Félix Desmeules-Trudel,
University of Toronto Mississauga,
Canada

*Correspondence:

Haley Weaver
hjweaver@wisc.edu

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 April 2021

Accepted: 23 August 2021

Published: 24 September 2021

Citation:

Bacon D, Weaver H and
Saffran J (2021) A Framework for
Online Experimenter-Moderated
Looking-Time Studies Assessing
Infants' Linguistic Knowledge.
Front. Psychol. 12:703839.
doi: 10.3389/fpsyg.2021.703839

Online data collection methods pose unique challenges and opportunities for infant researchers. Looking-time measures require relative timing precision to link eye-gaze behavior to stimulus presentation, particularly for tasks that require visual stimuli to be temporally linked to auditory stimuli, which may be disrupted when studies are delivered online. Concurrently, by widening potential geographic recruitment areas, online data collection may also provide an opportunity to diversify participant samples that are not possible given in-lab data collection. To date, there is limited information about these potential challenges and opportunities. In Study 1, twenty-one 23- to 26-month-olds participated in an experimenter-moderated looking-time paradigm that was administered via the video conferencing platform Zoom, attempting to recreate in-lab data collection using a looking-while-listening paradigm. Data collected virtually approximated results from in-lab samples of familiar word recognition, after minimal corrections to account for timing variability. We also found that the procedures were robust to a wide range of internet speeds, increasing the range of potential participants. However, despite the use of an online task, the participants in Study 1 were demographically unrepresentative, as typically observed with in-person studies in our geographic area. The potentially wider reach of online data collection methods presents an opportunity to recruit larger, more representative samples than those traditionally found in lab-based infant research, which is crucial for conducting generalizable human-subjects research. In Study 2, microtargeted Facebook advertisements for online studies were directed at two geographic locations that are comparable in population size but vary widely in demographic and socioeconomic factors. We successfully elicited sign-up responses from caregivers in neighborhoods that are far more diverse than the local University community in which we conduct our in-person studies. The current studies provide a framework for infancy researchers to conduct remote eye-gaze studies by identifying best practices for recruitment, design, and analysis. Moderated online data collection can provide considerable benefits to the diversification of infant research, with minimal impact on the timing precision and usability of the resultant data.

Keywords: infancy, online, eye-gaze, methodology, diversity, recruitment

INTRODUCTION

Developmental researchers face a multitude of barriers to completing research, particularly in determining methodologies appropriate for measuring various cognitive phenomena and participant recruitment. In particular, infants cannot provide verbal responses to interrogate underlying cognitive processes and thus researchers must rely on implicit behaviors such as eye-gaze. Moreover, infant samples are difficult to recruit and subject to high attrition rates (Enders, 2013; Klein-Radukic and Zmyj, 2015; Nicholson et al., 2015), resulting in local convenience sampling and limited generalizability. The coronavirus pandemic further complicated developmental work by limiting feasible in-person methodologies. The availability of well-defined virtual data collection methods for use with infants and young children is limited compared with adult methods, creating delays for research programs that rely on methods like eye-gaze. For many common infant cognition tasks, eye-gaze behavior is coded relative to audio stimuli presentation, which necessitates that the timing of stimulus presentation and data uptake must be quite accurate. Families may vary in the speed and reliability of their home internet connection, which can have downstream impacts on the timing of stimulus presentation and the frame rate of video recordings. In fact, prior research using videoconferencing describes internet connectivity, stability, and video quality as some of the disadvantages to collecting data virtually (Archibald et al., 2019). This may be particularly challenging for studies focused on language learning due to the need to integrate audio-visual stimuli. For example, to assess word recognition using eye-gaze behavior, researchers analyze visual attention to a particular image after hearing the onset of a word. Thus, inferences in these paradigms crucially depend on the temporal alignment of looking behavior to the onset of an auditory stimulus.

Existing online methods for developmental research are primarily geared toward either unmoderated data collection or older children. LookIt, an infant and child research platform based at MIT, allows researchers to upload unmoderated experiments to the platform to be completed by participants and caregivers (Scott and Schulz, 2017). TheChildLab is an experimenter-mediated video chat platform for study administration used with slightly older populations (aged 5+; Sheskin and Keil, 2018). Using TheChildLab, Sheskin and Keil (2018) were able to replicate in-lab effects for this age group. Both LookIt and TheChildLab demonstrate the feasibility of doing online research with developmental populations, though only a limited set of tasks have been verified for use *via* online platforms.

In addition to these platforms, developmental researchers have been using Zoom for experimenter-mediated studies. For example, Smith-Flores et al. (2021) used Zoom to replicate several in-lab findings using violation of expectation paradigms with 15-month-old. Importantly, Smith-Flores et al. (2021) reported global measures of looking-time (i.e., average proportion of looks), but did not examine moment-by-moment changes in visual attention in response to a stimulus, as is common with lab-based experiments focused on early language

development. Although Sheskin and Keil (2018) and Smith-Flores et al. (2021) suggest that experimenter-moderated data collection is promising, it is unclear how variability between participants' home set-ups impacts the subsequent data quality. Lack of internet access or poor internet connectivity could render participants' video data unusable due to inconsistent stimulus presentation. However, limiting participation to only high-speed internet users could create a significant barrier to families' ability to participate in online studies further perpetuating the issue of infant samples drawn predominantly from highly educated and wealthy families. Despite the unknown variability in home-set ups and internet access, the use of online data collection methods could provide an opportunity to ameliorate another persistent problem in developmental research: the lack of diversity of infant participants in lab-based studies.

Psychological research with human participants has historically relied on White, upper-, to upper-middle-class convenience samples. The resulting findings are representative of the participant group, but not necessarily of the wider, more diverse population the results are often applied to. Several research bodies have long recommended diversification of both researchers and participants in psychological research, placing the onus on the researchers to recruit members of underrepresented groups (National Institutes of Health, 1994; American Psychological Association, 2003; American Psychological Association, APA Task Force on Race and Ethnicity Guidelines in Psychology, 2019). Despite this push from respected institutions like the NIH and APA, psychological research has continued to primarily consist of Western, Educated, Industrialized, Rich, and Democratic (WEIRD; Henrich et al., 2010) convenience samples, tested at or in the immediate areas around universities. There has been a push for researchers to report their diversity (or lack thereof) in their research proposals and publications. Many proposals and publications report that their sample is representative of the local population; however, simply matching local census proportions does not make results generalizable. The geographic locations of universities and their surrounding population demographics place limitations on the population that can access in-person studies. The internet, and its increasingly pervasive presence in homes around the world, presents the opportunity to reach a more diverse participant sample.

Online recruitment and research with adult participants support the assertion that online data collection can lead to a diversification of participants. In adult studies, Amazon Mechanical Turk (MTurk) samples provide the most ethnic and socioeconomic (SES) diversity out of all of the adult study platforms (Casler et al., 2013), though it is not the only platform that works to recruit diverse samples (Casler et al., 2013; Buhrmester et al., 2016). Importantly, data quality was similar across participants regardless of whether they were recruited and participated *via* MTurk, social media, or face-to-face behavioral testing (Casler et al., 2013). These adult findings provide hints that variability in home data collection environments can have beneficial impacts on diversity without significant differences in experimental results. By extending recruitment efforts and increasing the diversity of participant

samples through online methods, results are more representative across race and SES.

Online recruitment methods, although able to reach a wider audience than typically is reached in the community surrounding universities, are not without their own impediments. Facebook, one of the most widely used digital recruitment platforms, is much more popular with White users, while Instagram is more popular with Latinx and Black users (Krogstad and Pew Research Center, 2015). The 2012 Facebook acquisition of Instagram, and the integration of the ad features on both platforms, allows ads that originate on one platform to appear in feeds of users on the other platform. As of August 2020, Facebook no longer allows demographic information pertaining to race to be used in targeted ads. Displaying the same ad across platforms may ameliorate disproportionate ad display to specific racial groups. Researchers working with adults have successfully increased the racial diversity of their samples without race-based microtargeting by targeting zip codes with larger non-White populations while keeping other targeted features constant (i.e., targeting people with particular sets of interests; Pechmann et al., 2020). While these approaches work well for adult samples (Casler et al., 2013; Pechmann et al., 2020), it is unclear whether a similar digital approach to recruitment and study administration is plausible for infant studies.

There is some evidence that online methods of recruitment targeting parents could be effective at recruiting more diverse infant and child participants for developmental research. Recruitment of parents *via* MTurk is fast, cheap, and results in more diversity than relying on Listservs (Buhrmester et al., 2016; Dworkin et al., 2016). Facebook ads targeting parents of specific races and ethnicities also yield more diversity than relying on Listservs or posting flyers around communities (Dworkin et al., 2016; Jang and Vorderstrasse, 2019). The LookIt platform provides a more racially diverse and representative United States participant sample than in-lab studies (Scott and Schulz, 2017). Although online data-sharing platforms, like the Databrary Project¹ and the Child Language Data Exchange System (CHILDES) allow researchers to access data from studies conducted globally (MacWhinney, 2000; Adolph et al., 2017), developmental researchers have called for greater efforts to conduct studies with representative samples. In particular, several scholars have proposed the development of Collaboration for Reproducible and Distributed Large-Scale Experiments (CRADLE) where there can be a combining of data from multiple data collection sources, addressing the need for more diverse and inclusive samples (Sheskin et al., 2020).

The present studies aim to address gaps in our understanding of online data collection and recruitment methods by (1) evaluating whether looking time data collection and retention with infants is possible across a range of home-set up variables and (2) investigating whether online data collection can facilitate more representative research by recruiting more diverse potential participants.

To determine whether real-time eye-gaze behavior can be captured in a Zoom study protocol, Study 1 includes a standard looking-while-listening (LWL) procedure with static images (Fernald et al., 2008). In face-to-face lab tasks, this method typically uses an eye-tracker to collect data, though there is evidence that hand-coding the data from video using custom software is not only reliable, but actually yields more usable trials and larger effect sizes than remote eye-tracker data (Venker et al., 2020). Study 1 was designed to test the hypothesis that experimenter-moderated studies over Zoom yield high quality infant data using LWL. The session recordings used for our primary method of data collection are the result of participant screen-sharing, which leaves the data subject to several uncontrolled variables: the frame rate of Zoom, the internet upload speed of the participant's computer, and the internet download speed of the experimenter's computer. Thus, a central goal of Study 1 is to evaluate these factors to determine whether variability in home set-up hinders interpretation of looking time data collected online. To assess the quality of the data, we examine the time course, average speed (reaction time, RT), and accuracy of looks to the targets during the LWL task. We compare these measures of data quality to a sample of data from Peekbank (Zettersten et al., 2021), an open-source database of in-lab LWL studies. We predict that experimenter-moderated Zoom LWL will approximate in-person data collection in timing precision and word recognition accuracy.

While Study 1 provided insight into the validity of online eye-gaze data across a range of set ups, the study did not adequately address diversity initiatives assumed to be improved using online data collection. Indeed, the participants in Study 1 were no more diverse than those we typically see for in-lab studies in our community. Thus, Study 2 was designed to determine if online recruitment targeting more diverse geographic locations could increase the diversity of participant signups for future study participation. Specifically, we asked whether more diverse families than those in the surrounding local population would express interest in participating in online experiments as a result of a microtargeted social media advertisement. We selected two locations in the same US state and matched them to be comparable in population size. Site 1 was predominantly Black and lower SES, and Site 2 was predominantly White and higher SES. We created a single ad using photos of lab participants and experimenters during a mediated online study session; the photos depicted racial diversity in both the participants and the experimenters (see **Figure 1**). Using Facebook's system for creating ads, we targeted the ad to the zip codes of the two sites, and then added interest-based targeting details that were race-neutral (e.g., parenting and childbirth). The ad linked to a lab sign-up page where the families of potential participants could enter information to be contacted for future studies. Based on prior research suggesting the efficacy of using diverse targeted advertisements for recruitment, we predicted that we would obtain more diverse participant sign-ups when an ad is targeted to people living in a more diverse area (Dworkin et al., 2016; Jang and Vorderstrasse, 2019; Pechmann et al., 2020).

¹The Databrary Project is accessible to institutionally affiliated researchers at databrary.org



FIGURE 1 | Facebook advertisement used for microtargeting.

STUDY 1: ONLINE LOOKING TIME STUDY

Methods

Infants saw a pair of familiar objects on each trial and heard the speaker ask for a target item. We predicted that infant word recognition and performance on the LWL task would be comparable to the results of lab studies that use familiar nouns as stimuli accessible on Peekbank (Zettersten et al., 2021), a database of LWL studies. Thus, we expected infants to show an increase in target looking following the onset of the noun. However, note that in the current paper, we will focus on assessing data quality by testing the hypothesis that data collected *via* Zoom will approximate an in-lab sample from Peekbank.

Participants

Twenty-one full-term, monolingual English-learning infants (nine females) with a mean age of 25.0 months (23.0–26.0) were included in the analyses. Families were recruited from an existing research

database tied to the local community ($n=19$ identified as non-Hispanic White; $n=1$ identified as multiracial; and $n=1$ identified as Hispanic White). Caregivers reported that their children had no history of developmental concerns, heard fewer than 10h per week of another language, and were currently free of ear infections. Eight additional participants were excluded due to: technical error ($n=4$), experimenter error ($n=2$), or failure to complete the task ($n=2$). Caregivers provided written informed consent. All experimental protocols were approved by the local Institutional Review Board. Data were collected between 10/20 and 02/21 as part of a larger project investigating the relation between words and knowledge of object functions.²

Materials

A female native speaker of English recorded 12 sentences using infant directed speech. Each sentence included one of four

²The OSF repository containing the stimuli for the larger project can be found at https://osf.io/nuecw/?view_only=0208ac75881148c9b602e22520a1bdc1

carrier phrases (i.e., “Find the [target noun]!,” “Look at the [target noun]!”) followed by a target noun (*apple*, *ball*, *crayon*, and *toothbrush*). Still images of the target nouns were selected (Brodeur et al., 2014) and placed on a grey 360 × 360-pixel gray background using GIMP.³ Three unique images were chosen for each of the target nouns for a total of 12 still images. Each object occurred equally as often as a target and distractor.

Procedure

Participants were tested *via* the videoconferencing platform Zoom. Caregivers completed a home setup procedure guided by an experimenter to maximize their lighting, screen display, and child positioning. Caregivers used their own computer (laptop or desktop) to access a personalized study link and shared their screen with the experimenter for screen recording. Study participation was limited to those with access to a computer due to the inability to screenshare and the constraints on stimuli size when using tablets or smartphones. Caregivers closed their eyes during testing to minimize bias. Each Zoom session, including the caregiver’s shared screen displaying the experimental procedure, was recorded locally by the experimenter for offline eye-gaze coding, frame-by-frame (40ms), using an open-source program for eye-gaze coding (Peyecoder; Olson et al., 2020).

Infants’ real-time word comprehension was assessed using LWL (Fernald et al., 2008). On each trial, two pictures of familiar objects were displayed simultaneously in silence for 1,000ms. Stimuli were aligned horizontally at a fixed distance of 540 pixels, which was held constant across all participants regardless of screen size. Infants heard speech labeling one of the objects in a carrier phrase (767ms) ending in the target noun (708ms). Infants were allowed to view the images for 2,025ms after the offset of the target noun for a total trial length of 4,500ms. There were six test trials for each target noun for a total 24 test trials. Trials were presented in a pseudorandom order in blocks of six interspersed with attention getters.

Internet Speed Test

To evaluate the impact of internet variability, we simulated the participant and researcher experience with the online task under different internet speeds using the developer tools in Google Chrome [Version 90.0.4430.85 (Official Build; x86_64)]. We tested four different internet speeds (2G, 3G slow, 3G fast, and 5G no throttling) to verify that events occurring on the Zoom recording reflect the events that a participant experienced. The internet tests were used to ensure that the events captured within our Zoom recordings can reliably be time-locked to the participant’s eye-movements. They also provided independent verification that we could include data from a range of home set-ups and did not need to exclude participants on the basis of internet speed.

For each internet speed test, two researchers imitated the experimental procedure by deploying the task in Google Chrome while participating in a Zoom call. One of the researchers, serving as the “participant,” shared their screen. Two videos were recorded

from each speed simulation. One video was recorded from the experimenter perspective using Zoom to imitate the data collected during an experimental session. The second video was a screen recording [QuickTime Version 10.5 (1015.2.1); recorded at 60fps] of the Google Chrome window running the experimental procedure from the participant perspective to capture a participant’s experience of the task at the current internet speed. A trained research assistant coded the trial onsets and offsets of the videos using Peyecoder (Olson et al., 2020).

Peekbank Data

In order to have a reasonable point of in-lab comparison for our experimenter-moderated online LWL task, we consulted Peekbank, a new open-source database of LWL studies (Zettersten et al., 2021). Using peekbankr, we searched for experiments testing infants in our target age range (23- to 26-month) whose primary language is English. We then filtered this sample for data testing familiar words (rather than nonce words). This yielded a sample of data from 126 participants across six studies (Yurovsky et al., 2013; Mahr et al., 2015; Frank et al., 2016; Yurovsky and Frank, 2017; Potter and Lew-Williams, in prep; Yurovsky et al., under review). One study was excluded for using a tablet-based LWL paradigm, which does not reflect our typical in-lab data collection paradigm using an eye-tracker. We also filtered the sample, limiting it to our specific target words (*apple*, *ball*, *crayon*, and *toothbrush*) of which only *apple* and *ball* were included in the dataset. Across three experiments (Mahr et al., 2015; Potter and Lew-Williams, in prep; Yurovsky et al., 2013), data from 70 participants for these two target words were obtained.

Coding

Trained research assistants coded eye movements frame-by-frame at a frame rate of 25fps using Peyecoder (Olson et al., 2020). Coders indicated whether infants were looking left, right, or off (i.e., in a gaze shift between images or looking off screen). Twenty-five percent of the videos were randomly selected and independently recoded. We evaluated reliability on three measures: (1) the percentage of gaze shifts that occur within a one-frame threshold (i.e., do coders agree on the timing of coded events?; shift agreement; 93.48%); (2) the percentage of event frames that have the same response between coders (i.e., do coders agree whether a frame is coded as left, right, or off?; frame agreement; 95.52%); and (3) the percentage of trials that have the same number of coded events between coders, impacting how many trials were used to calculate shift agreement (comparable trials; 85.68%).

Results

Internet Speed Tests

For each internet speed (2G, 3G slow, 3G fast, and 5G no throttling), we calculated the total number of comparable trials and the frame agreement between the two videos to assess whether the number of trials captured by the Zoom recording differed from the participant’s experience of the study. If there is an internet lag, the number of trials seen in the participant view could differ from the experimenter view. **Table 1** provides the

³GIMP is an open-source editing software and is accessible at <https://www.gimp.org/>

TABLE 1 | Internet speed simulation.

| | 2G | 3G slow | 3G fast | 5G no throttling |
|---------------------------|--------|---------|---------|------------------|
| Frame agreement | 100.0% | 100% | 100% | 100% |
| Number of trials | 35:35* | 36:36 | 36:36 | 36:36 |
| Participant: Experimenter | | | | |
| Trial response agreement | 100.0% | 100% | 100% | 100% |

The asterisk (*) denotes a different number of trials than the experiment total, illustrating trial loss.

frame agreement, number of trials, and trial response agreement between the experimenter and participant perspective videos. Regardless of internet speed, the experimenter and the participant videos aligned. When independently coded, the participant and experimenter videos at each speed level have the same number of frames, number of trials, and the same trial responses. Most importantly, the lack of frame disagreement suggests that internet speed is not a significant barrier for online participation in tasks using audio, images, and videos. Although slower internet speeds influenced the presentation of the experiment (Table 1), the recorded Zoom data capture this inconsistency, which allow for unrepresented trials to be skipped during data analyses. The present results suggest that high-speed internet is not a prerequisite for usable data quality in online studies.

Time Course of Looking Behavior

The aim of the looking time results is to introduce a new procedure for correcting eye-gaze data given variable frame rates, and to provide evidence that eye-gaze behavior timing information recorded online is interpretable and comparable to in-lab research gathered from Peekbank. We report data visualizations in the form of time course plots to visually assess whether the data approximate what would be expected from data collected in the lab. In particular, we plot the proportion of looks to the target as a function of time with confidence bands reflecting SE of the point estimate. No inferential statistics were conducted.

Infant looking behavior was coded frame-by-frame resulting in eye-gaze data every 40ms over the course of a trial. We computed the proportion of looks to the target visual stimulus [accuracy; looks to target/(looks to target+looks to distractor)] at each 40ms time bin averaging across trials and participants. We were interested in looks beginning at 300ms after the onset of the target word and ending 1,800ms after target word onset (Fernald et al., 2008). The target window was selected to reflect similar window of analyses used in prior LWL studies with toddlers (Swingley and Aslin, 2000; Swingley and Aslin, 2002; Fernald et al., 2008; Bergelson and Swingley, 2012; Zettersten et al., 2021). We excluded trials that did not include looks to either the target or distractor image for at least 50% of the frames. Across all participants, only 59 trials out of the total 504 trials were excluded using this criterion (i.e., 88% of trials were usable). Individual infants

contributed an average of 21 useable trials (range: 15–24) out of a maximum 24 trials in the study, with no infant contributing fewer than 50% of all trials.

To evaluate the reliability of timing data derived from Zoom recordings, we examined whether the number of frames per trial replicated the expected total number of frames given the length of a trial. A coded LWL trial was 3,900ms and therefore should include 98 frames of looking data (3,900ms/40ms per frame). For each infant, we calculated the number of frames recorded for each LWL trial. On average, there were 92 frames per trial (range=20–136), which is six frames less than expected given the trial length. Therefore, the average time elapsed per frames is longer (43.86ms) than the assumed 40ms frame rate of Zoom recordings. There is an inverse relationship between ms per frame and the number of frames in a trial, such that longer frame lengths indicate a fewer total number of frames on a given trial. This timing discrepancy has implications for data coding. In particular, if the onset of a target noun is expected to occur at frame 29 (1,167ms onset time/40ms per frame), then it is actually occurring at frame 27 on average (1,167ms onset time/43.86ms per frame) due to the longer average length of a frame. Furthermore, the difference in the number of ms per frame can vary from trial to trial for a given participant, with the length of a frame ranging from 28.69 to 195ms. Therefore, for some trials, the onset of the target word occurs later in the trial, at frame 40, while for others it could occur as early as frame 6.

Given the discrepancy between expected and actual frame rates, we plotted the time course of target looks using two different measures of time: (A) uncorrected time using the Peyecoder frame rate (40ms) and (B) corrected time using each infant's average frame rate. For each infant, we computed the frame rate for each trial by calculating the average number of ms that elapsed per frame (i.e., a frame rate of 25fps indicates that 40ms elapses per frame). We calculated the length of a frame (in ms) by dividing the total length of a trial (3,900ms) by the total number of frames within each trial. Each child's timing data were adjusted frame-by-frame using their by-trial frame rate. For example, an event occurring at frame 29 in the assumed frame rate of 40ms per frame (based on the Peyecoder output) would be adjusted to occur at frame 31 for an infant who had an actual frame rate of 37ms per frame. To normalize the data across participants, we calculated the mean frame rate by averaging across all trials contributed by all participants. The adjusted timing data were binned into 43.86ms increments (22.80fps) to have comparable time bins across infants. Thus, a looking event that occurred at an assumed 40ms was adjusted to occur at an actual 43.86ms. Normalizing the timing data in this manner results in 90 frames that increment in 43.86ms time windows from 0 to 3,900ms. This process ensured that trials with different frame rates could still be averaged together to yield group level looking accuracy across infants and across trials.

The results of the adjusted time course of looks, collapsed across participants, can be seen in Figure 2. Notably, infants increased the proportion of looks to the target item during the critical window from 300 to 1,800ms in both plots, suggesting that infants recognized the target words. However, the corrected

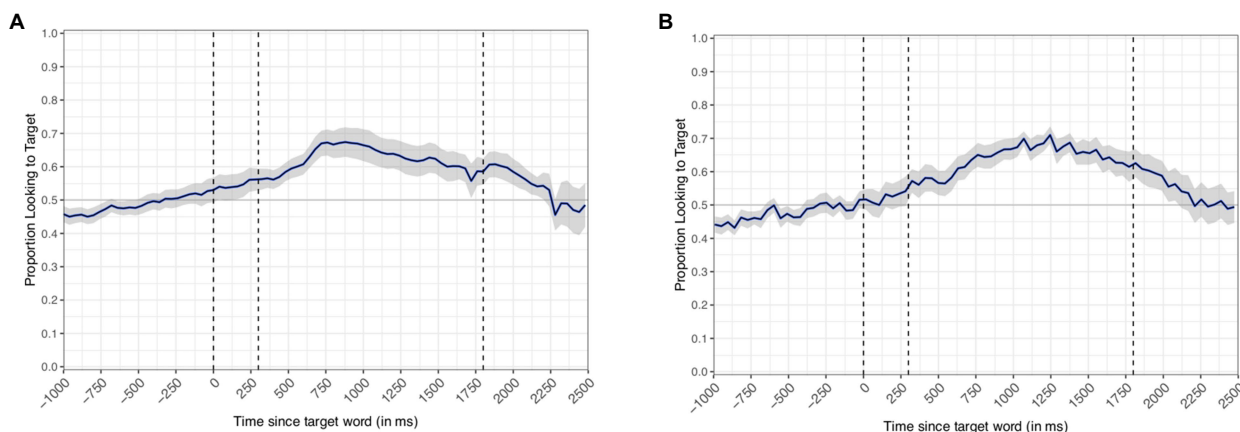


FIGURE 2 | Proportion of looks to a target image as a function of time. Dashed lines represent the onset of the target word (0ms) and demarcate the primary window of interest from 300 to 1,800ms. Graph **(A)** plots uncorrected time using the expected frame rate of 40ms per frame (25 fps). Graph **(B)** plots time that has been corrected to reflect each individual infant's mean-adjusted frame rate and normalized across participants to a frame rate of 43.84 ms per frame. Confidence bands represent SEM.

time course plot shown in **Figure 2B** demonstrates that when adjusting individual participants' data timing into the same time bins (using averaged frame rate) and then collapsing across participants, the looking behavior shows a looking pattern more similar to in lab eye-gaze assessment. Specifically, the accuracy in looks to the target in **Figure 2B** begins to diverge from chance closer to 300 ms (i.e., the approximate time it takes to execute a planned eye movement based on phonological information; for discussion of latencies to shift in LWL see Swingley and Aslin, 2000; Fernald et al., 2008; Zettersten et al., 2021) after the onset of the target noun as compared to **Figure 2A**, in which accuracy differs from chance beginning at approximately 150 ms. Further, **Figure 2A** reflects what would be expected by plotting timing data that results from a slower frame rate because the target word onset actually occurs in an early frame. The results of the time course plots suggest that online data collection can replicate previous findings for familiar word recognition (Fernald et al., 1998; Bergelson and Swingley, 2012; Bergelson and Aslin, 2017) despite some limitations due to variable frame rates.

The time course of target looks in the uncorrected data suggests that infants are shifting earlier than what would be expected in response to the auditory stimulus. We thus wanted to determine whether our process for timing corrections would more closely approximate the shift latencies in a sample of data from similar tasks collected in-person. Thus, we assessed whether there were significantly later shifts to the target in LWL data collected in person compared to the current sample of data collected *via* Zoom. Based on the time course of looks seen in **Figure 2**, we defined a window of analysis from –300 to 200 ms. This analysis window reflects a period of time when the average curve in the uncorrected timing data begins to deviate from chance to a point in time when the confidence bands do not include chance responding (**Figure 2**). Importantly, this time period occurs earlier than we would normally expect to see eye gaze behavior in response to the spoken words

with most studies approximating looking behavior to begin around 300 ms (e.g., Fernald et al., 1998; Swingley and Aslin, 2002; Fernald et al., 2008; Garrison et al., 2020). We expected that in-person LWL data, as represented by a Peekbank sample, would have significantly later shifts to the target than the uncorrected Zoom LWL data. However, if our adjusted time course is a more veridical representation of the task, we would expect that the in-person LWL data would not differ from the corrected Zoom LWL data. To test this hypothesis, we identified all trials in which an infant was fixating the distractor at the onset of the analysis window. We then calculated the latency to shift to the target image. We fit a linear mixed effects model (LMEM) predicting shift latency from the different datasets (i.e., uncorrected Zoom data, corrected Zoom data, and Peekbank data) including a by-subject random intercept and a by-item random intercept. We coded Peekbank as the reference group to compare whether the data collected *via* Zoom differed significantly from data collected using LWL in-lab. The average latency to shift was significantly later ($M=155.556$) in the Peekbank dataset compared to the uncorrected LWL data collected *via* Zoom [$M=134.460$; $b=-34.997$; $t(1, 39.697)=-2.224$; $p=0.032$; 95% CI $(-65.844, -4.151)$]. There was no significant difference in timing between the Peekbank dataset and the corrected LWL Zoom data [$M=141.074$; $b=-16.069$; $t(1, 52.328)=-0.939$; $p=0.352$; 95% CI $(-49.601, 17.464)$; **Figure 3**]. This analysis supports our contention that the timing correction serves an important data preprocessing step in adjusting the timing of the trial so that it more accurately reflects the actual presentation of stimuli.

Reaction Time Results

We were interested in evaluating how the timing of shifts in response to an auditory stimulus in the present Zoom study compares to an in-person LWL designs. To examine this question, we compared the average reaction time (RT) for the uncorrected and corrected timing data to samples of LWL data from Peekbank

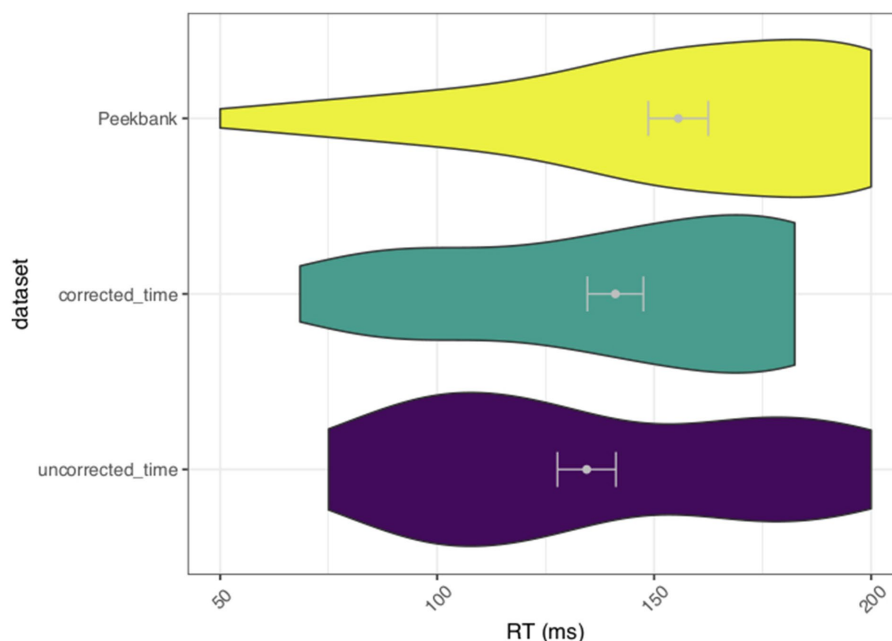


FIGURE 3 | Average latency to shift to the target from distractor during a window from -300 to 200 between Peekbank data and the Study 1 corrected and uncorrected timing data. Error bars represent SEM.

(Zettersten et al., 2021) that test a subset of the target words (i.e., apple and ball) in our target age range (23- to 26-month). RTs were defined as the average time it takes an infant to shift from the distractor image to the target image on a given trial from the onset of the target word (0ms) to 1,800ms post-word onset (Fernald et al., 2008). We would expect that the RTs calculated using the corrected timing data should be more similar to samples drawn from Peekbank than the uncorrected timing data.

Reaction time was calculated for all trials in which an infant was initially fixating the distractor at the onset of the target time window. For each trial, we calculated an infant's latency of the first shift to the target image from the distractor image (Fernald et al., 2008). We then filtered out RTs that were later the predetermined window length. This definition of RT does not include time to fixate the target, but rather demonstrates the time it takes to process the auditory stimulus and make a behavioral response.

We fit a LMEM to compare RTs within the target window (0–1,800 ms) to assess average shift latency in response to the target words. We regressed RTs on a variable for the different datasets including a by-subject random intercept and slope for dataset and a by-item random intercept and slope for dataset. Dataset was contrast coded using Peekbank as the reference group to assess whether the average RTs from the corrected and uncorrected Zoom data differ significantly from RTs typically seen in-lab. RTs from the Peekbank dataset were significantly longer ($M=909.420$) than both the corrected [$M=417.747$; $b=-550.366$; $t(1, 6.357)=-7.525$, $p<0.05$; 95% CI $(-693.723, -407.010)$] and uncorrected online data [$M=399.722$; $b=-571.914$; $t(1, 5.276)=8.067$; $p<0.05$; 95% CI $(-710.866, -432.963)$]. It is possible that the methodological

differences between in-lab and online studies (i.e., screen size, distance from the monitor, and number of test trials per word) could account for faster RTs in the online experiment. We return to this possibility in the Study 1 Discussion.

Accuracy Results

We were interested in whether the data collected *via* Zoom would approximate word recognition accuracy for familiar words that is expected from in-lab LWL designs. Thus, we compared the average proportion of looks to the target image (accuracy) in the corrected online timing data to the sample of LWL data from Peekbank. For each dataset, we computed an infant's by-trial average accuracy during the window from 300 to 1,800ms (Fernald et al., 2008). If the mode of data collection has minimal impact on the resultant data quality, we would expect minimal differences in accuracy across these study types.

We fit a LMEM regressing accuracy on dataset type including a by-subject random intercept and a by-item random intercept. We also included an offset at 0.5 to evaluate whether average accuracy differed significantly from chance responding. For this analysis, the corrected online dataset served as the comparison group (coded as 0) to determine whether the datasets derived from in-lab studies (Mahr et al., 2015; Potter and Lew-Williams, in prep; Yurovsky et al., 2013; Yurovsky and Frank, 2017; Yurovsky et al., under review) differed significantly in average accuracy for familiar word recognition from the current study. We report Holm-Bonferroni corrected values of p to account for multiple comparisons (Holm, 1979).

On average, infants' accuracy on the online LWL task was significantly greater than chance [$b=0.162$; $t(1, 76.570)=4.664$; $p<0.05$]. Accuracy on the online LWL task was significantly

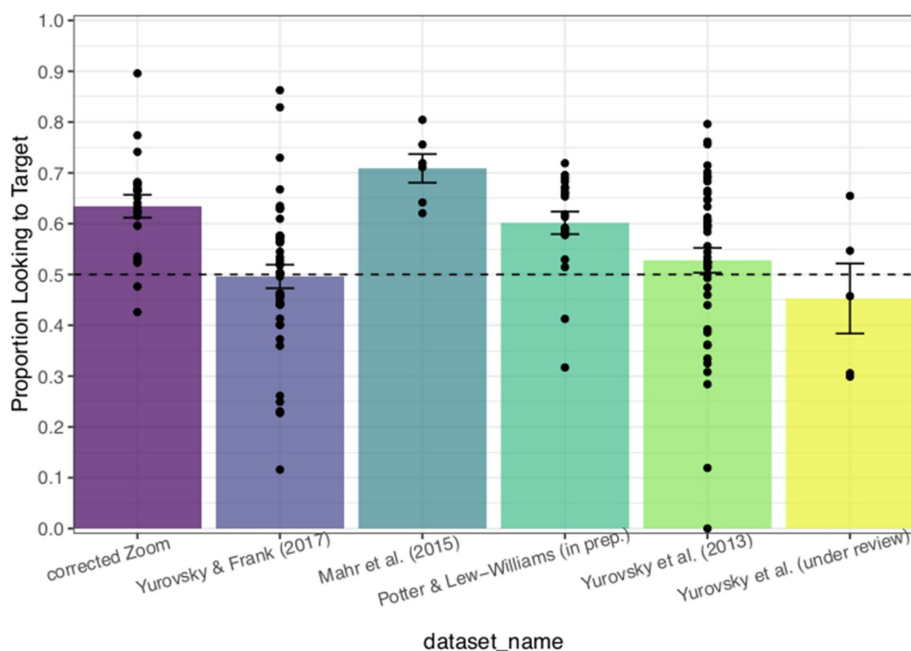


FIGURE 4 | Proportion of looks to the target (accuracy) by study, comparing Study 1 to selected Peekbank studies. Chance looking behavior is denoted by the dashed line at 0.5 on the y-axis. Error bars represent SEM.

different from accuracy in the in-lab data collected in Yurovsky and Frank (2017) [$b = -0.151$; $t(1, 165.880) = -3.417$; $p < 0.003$], Yurovsky et al. (2013) [$b = -0.160$; $t(1, 237.895) = -3.990$; $p < 0.05$], and Yurovsky et al. (under review) [$b = -0.222$; $t(1, 296.866) = -2.900$; $p = 0.012$]. As can be seen in **Figure 4** accuracy on the online task is significantly greater ($M = 0.634$) than accuracy on the three in-lab tasks ($M = 0.496$, $M = 0.528$, and $M = 0.453$, respectively). The range of in-lab familiar word recognition accuracy in **Figure 4** suggests that the data collected *via* Zoom is feasible and valid.

Study 1 Discussion

In Study 1, we evaluated the timing precision of online experimenter-moderated eye-gaze measures of LWL. Approximately 88% of the trials in the were usable, which constitutes similar rates of data loss to in-person data collection (Wass et al., 2013; Venker et al., 2020; but also see Oakes, 2010 for a call for greater transparency in reporting eye tracking data). Data quality was not significantly impacted across a range of different internet speeds, suggesting that various levels of internet connectivity can be supported in online data collection using these paradigms. Although internet connectivity did not preclude participation, it did contribute to immense frame rate variability across participant video recordings. Individual variability in timing can be corrected during data preprocessing using group-level average frame rates. Using this correction technique, we can account for differences in home testing conditions that are not typically seen in the lab that utilizes a single set of technical equipment. Taken together, these results suggest that looking-time behaviors can be captured *via* videoconferencing across a variety of home-set ups.

Across participants, there were differences in RT; however, the corrected and uncorrected RTs from the present sample

were more similar to one another than they were to the in-lab RTs in the Peekbank data. This may be due to the differences in set-up for the online study administration and in-lab study administration. In Study 1, participants were seated in their caregivers' lap like they would be in the lab, but were watching the study visuals occur on a much smaller screen and at a much closer distance than they would in-lab. The current design also tested each target word six times and the same images were seen multiple times, which may account for faster processing speeds.

In sum, the findings from Study 1 suggest that online data collection is feasible and yields high quality data, particularly when the data are adjusted to reflect frame rates. Experimenter-moderated online studies may be a way to collect more equitable and representative samples given that access to high-speed internet is not a requirement for participation. Families, who typically would not be able to attend in lab sessions due to scheduling, travel, or other barriers could join a 20 min Zoom session to participate during a time and location that is convenient for them. Despite this theoretical benefit to online testing, Study 1 included a homogenous, predominantly White sample. Importantly, simply moving to a virtual platform did not ameliorate the issue of diversity in the participant sample. In Study 2, we test a recruitment method to increase the diversity of our participant pool for future online studies.

STUDY 2: ONLINE RECRUITMENT

Study 1 provided preliminary promise that data collection using infant looking time measures was possible across several different home set-up variables. These results suggest that internet

connectivity should not preclude participation in our experiment. Yet, the sample in Study 1 was homogenous and WEIRD despite being conducted online. Thus, simply administering the task in the home was not sufficient to recruit a more diverse, representative sample. This was likely due to the use of our existing participant database, developed for the local community, for Study 1. In Study 2, we examined whether targeting our recruitment efforts to locations with more diverse populations could provide a more representative sample for future online studies. At present, it is unknown whether ads targeting more diverse locations actually lead to more diverse potential research participants in developmental studies. The goal of this study was to determine whether microtargeting based on location can alter the demographics of respondents to our research advertisements.

Methods

Participants

This study focused on caregivers who responded to Facebook advertisements that were microtargeted to display in two different cities. Upon clicking a sign-up link in the Facebook ad, caregivers voluntarily provided contact information that can be used to alert the family of future study opportunities. For analytic purposes, we considered a respondent to be a unique sign-up that included a child's name and caregiver contact information. Fifty-one respondents were included in the data analyses (Site 1 $N=14$) for children that ranged in age from *in utero* and expected to be born in 2021 to 99-month. Eight additional respondents who signed up were excluded because they did not provide contact information. Note that we only collected information about names, ages, mailing addresses, and other contact information from respondents; we did not have IRB approval to collect any demographic information (e.g., race, SES) on our sign-up link. Thus, as noted below, we used census-tract data as a proxy to estimate demographic information about the caregivers who responded to our ad.

Materials

The ad featured two photos of participants and experimenters during study administration (**Figure 1**) (1) featuring two experimenters (one White, one Black) and a caregiver/child duo (both Black) and (2) featuring one experimenter (multiracial) and a caregiver/child duo (both White). These photos were selected based on work indicating that diversity in advertisements begets diversity in recruitment (Avery et al., 2004; Walker et al., 2012; Pechmann et al., 2020).

Procedure

This study focused on caregivers who responded to microtargeted Facebook advertisements that were directed at zip codes in two Midwestern cities in the same state with different demographic profiles. Site 1 has a predominantly Black population and Site 2, the catchment area for our in-lab studies, has a predominantly White population. We targeted the ad to a subset of zip codes within each city to ensure that the recruitment catchment areas were comparable in population size but varied on other key demographic features related to diversity (i.e., household income; see **Table 2**).

TABLE 2 | Demographic and socioeconomic factors for the two target recruitment sites (set of zip codes targeted in the Facebook ad) as reported on the American Community Survey (ACS) 5-year estimates.

| | Site 1 | Site 2 |
|---|-------------|-------------|
| Population | 31,565.83 | 34,705.83 |
| Race and ethnicity | | |
| Hispanic | 5.00% | 10.00% |
| American Indian or Alaskan Native | 1.00% | 1.00% |
| Asian | 6.00% | 10.00% |
| Black or African American | 68.00% | 10.00% |
| Native Hawaiian or Pacific Islander | 0.00% | 0.00% |
| Other | 3.00% | 4.00% |
| White | 26.00% | 79.00% |
| Socioeconomic factors | | |
| Cost of living index | 97.18 | 98.35 |
| Median <i>per capita</i> income | \$39,479.83 | \$61,326.83 |
| Percentage of children below the poverty line | 39.00% | 16.00% |

These locations vary on racial, ethnic, and socioeconomic (SES), but are comparable in population size.

The ad targeted users aged 18–65+ with interests matching some or all of the following: family, motherhood, fatherhood, parenting, breastfeeding, childbirth, day care or early childhood education, job titles that included “science,” parents: new parents (0–12 months), or parents with toddlers (01–02 years).

Collecting Demographic Variables

Study 2 primarily focused on recruiting participants for future participation in online studies. The current respondents did not partake in any research. Thus, neither did they provide consent, nor did they contribute any data or demographic information. Upon sign-up, caregivers voluntarily provided contact information to be used to alert the family of their child's eligibility to participate in a study. Facebook does not currently provide ad users with demographic features (other than age) about those who interact with their advertisement engagements. Thus, to assess the success of our ad in eliciting responses from more diverse populations, we relied on demographic metrics drawn from the American Community Survey (ACS) 5-year estimates (United States Census Bureau, 2020). We identified the US census tract for the home address of each sign-up we received. For each respondent, we use the demographic features (i.e., race, ethnicity, income, etc.) available for their *census tract* as a proxy for the likely demographic features of the participant. A US census tract accounts for one square mile of a geographic location. Thus, for Site 1, each percentage estimate reflects the proportion of people out of 6,188 people within a given area that identify with the demographic feature of interest, while Site 2 reflects the proportion of people out of 3,037 people within a given area (United States Census Bureau, 2020). For example, if 95% of people within a census tract within Site 1 identify as Black or African American this can be interpreted as approximately 5,879 out of 6,188 people within the census tract identify as Black or African American. We acknowledge that these data

TABLE 3 | Facebook ad reach metrics.

| Facebook metric | Estimated value |
|-----------------|-----------------|
| Total reach | 13,392 |
| Percent women | 88.9% |
| Engagements | 577 |
| Reactions | 215 |
| Link clicks | 159 |
| Shares | 14 |
| Comments | 5 |
| Sign-ups | 59 |

may not accurately represent the demographic characteristics of our individual respondents. However, the ACS demographic estimates allow us to empirically evaluate whether microtargeted Facebook ads resulted in sign-ups from groups of participants located in areas that are more diverse than those typically targeted for research participation.

Results

Facebook Ad Results

Facebook estimated that our microtargeted advertisement (\$400.00 USD total for a 14-day run) reached 13,392 people and that there were 577 interactions with the advertisement (see **Table 3**). Facebook defines interactions to include shares, likes, comments, and clicks. Of these interactions, 159 of them were clicks on links included in the ad (lab website link and sign-up link), which resulted in 59 new participant sign-ups. Metrics revealed that the ad was primarily presented in FB mobile app feeds (12,956 people out of 13,392 total reach).

Demographics by Targeted Site Location

The aim of the microtargeted Facebook ad was to provide a more diverse pool of participants than typically generated by local convenience sampling. Because we did not have direct information about the demographics of our sample (as noted earlier, these data represent sign-ups for future studies rather than consented participants), we estimated the demographics of the respondents using census tract-level data. We identified the census tract number for each unique address provided at sign-up which resulted in 35 unique tracts. Two additional tracts were excluded for being located out of the target state. Each census tract was then coded as located in either Site 1 (13 tracts) or Site 2 (22 tracts). To evaluate the diversity of each group of potential participants (Site 1 vs. Site 2), we queried the ACS 5-year estimates for four factors related to diversity: racial makeup, ethnic background, educational attainment for the population over 25-years-old, and median household income for each of the census tracts. To determine whether there were potential differences in the demographics of respondents from each site, we ran linear regression models using the *lmSupport* package (Curtin, 2018) in R (Version 1.2.1335; R Core Team, 2019). We report the results of the regression analyses for each diversity metric, separately.

Racial Makeup

We evaluated the potential racial diversity of our respondents by comparing the racial makeup according to the ACS estimates

derived from respondents' census tracts between the two site locations. We predicted that a higher proportion of respondents from census tracts located in Site 1 would belong to more diverse racial categories than those from Site 2, where we expected that the majority of the respondents would identify as White. To test this hypothesis, we computed proportion of the population (number of people within a respondent's census tract that identify as a racial category/total population within the census tract) that identified as American Indian or Alaskan Native, Asian, Black or African American, Native Hawaiian or Pacific Islander, White, or Other. Thus, for each respondent we calculated six proportions, corresponding to each of the racial categories from their census tract data. We fit a linear model regressing these proportions on race (dummy coded with White as the reference group), site (centered, coded Site 1 = -0.5 and Site 2 = 0.5), and their interaction. The results of the linear model are reported in **Table 4**. Given that the variable race was dummy coded, each estimate indicates whether there is a significant difference in the average proportion of White people compared to each of the other racial categories (i.e., proportion of White people compared to the proportion of Black people). All values of *p* were corrected for multiple comparisons using the Holm-Bonferroni approach (Holm, 1979).

Overall, a significantly higher proportion of residents across both sites identified as White ($M=0.590$) as compared to Black or African American ($M=0.27$), Asian ($M=0.070$), or American Indian or Alaskan Native ($M=0.010$). Importantly, there was also a significant race by site interaction [$F(5, 288)=80.704$, $p<0.05$, $\eta^2p=0.584$]. This significant interaction indicates that the mean proportion of White people compared to the proportion of people that identified as each of the other racial categories differed depending on Site location (**Figure 5**). Specifically, compared to Site 2, respondents whose census tracts were located in Site 1 had a significantly higher proportion of people that identified as Black [$b=-0.967$, $F(1, 288)=402.11$, $p<0.05$], Asian [$b=-0.434$, $F(1, 288)=81.06$, $p<0.05$], American Indian or Alaskan Native [$b=-0.469$, $F(1, 288)=94.76$, $p<0.05$], Native Hawaiian or Pacific Islander [$b=-0.472$, $F(1, 288)=95.65$, $p<0.05$], or Other [$b=-0.471$, $F(1, 288)=95.46$, $p<0.05$; **Table 4**]. The Facebook ad specifically targeted a predominantly Black location (Site 1) and a predominantly White location (Site 2). The results from the regression analysis suggest that the respondents from Site 1 represent a more diverse set of potential participants. As can be seen in **Figure 5**, respondents' census tracts in Site 1 included a higher proportion of Black or African American people ($M=0.627$) compared to White people ($M=0.252$), while respondents' census tracts in Site 2 included a smaller proportion of Black or African American people ($M=0.130$) compared to White people ($M=0.722$).

Ethnic Background

We examined ethnic diversity between the two targeted sites by conducting a parallel analysis to the analysis of racial diversity. We calculated the proportion of the population that identified as Hispanic or non-Hispanic for each of the census tracts (number of people within a respondent's census tract that identify as Hispanic or non-Hispanic/total population within the census tract). Given the target site demographics (**Table 2**), we predicted that Site 2 would have a slightly higher

TABLE 4 | Results of regression analyses by demographic metric.

| Demographic metric | <i>b</i> | <i>F</i> | <i>p</i> | <i>df</i> | <i>R</i> ² |
|--|----------|----------|----------|-----------|-----------------------|
| <i>Race</i> | | | | 288 | 0.845 |
| Site | 0.469 | 189.64 | 0.000 | | |
| Asian vs. White | −0.204 | 24.76 | 0.000 | | |
| Black or African American vs. White | 0.375 | 84.03 | 0.000 | | |
| American Indian or Alaskan Native vs. White | −0.246 | 36.09 | 0.000 | | |
| Native Hawaiian or Pacific Islander vs. White | −0.250 | 37.27 | 0.000 | | |
| Other vs. White | −0.231 | 31.78 | 0.000 | | |
| Asian vs. White * Site | −0.434 | 81.06 | 0.000 | | |
| Black or African American vs. White * Site | −0.967 | 402.11 | 0.000 | | |
| American Indian or Alaskan Native vs. White * Site | −0.469 | 94.76 | 0.000 | | |
| Native Hawaiian or Pacific Islander vs. White * Site | −0.472 | 95.65 | 0.000 | | |
| Other vs. White * Site | −0.471 | 95.46 | 0.000 | | |
| <i>Ethnicity</i> | | | | 78 | 0.969 |
| Site | 0.000 | 0.000 | 1.000 | | |
| Ethnicity | −8.965 | 903.474 | 0.000 | | |
| Site * Ethnicity | 0.061 | 2.723 | 0.103 | | |
| <i>Education</i> | | | | 240 | 0.476 |
| Site | 0.135 | 19.025 | 0.000 | | |
| Associate's degree vs. Graduate degree | 0.001 | 0 | 1.000 | | |
| Bachelor's degree vs. Graduate degree | 0.050 | 1.833 | 0.531 | | |
| High school diploma or less vs. Graduate degree | 0.358 | 92.991 | 0.000 | | |
| Some college vs. Graduate degree | 0.188 | 25.522 | 0.000 | | |
| Associate's degree vs. Graduate degree * Site | −0.123 | 7.854 | 0.022 | | |
| Bachelor's degree vs. Graduate degree * Site | 0.026 | 0.344 | 1.000 | | |
| High school diploma or less vs. Graduate degree * Site | −0.347 | 62.699 | 0.000 | | |
| Some college vs. Graduate degree * Site | −0.232 | 27.952 | 0.000 | | |
| <i>Income</i> | | | | 48 | 0.227 |
| Site | 23,428 | 14.13 | 0.000 | | |

proportion of Hispanic respondents than Site 1. We tested this hypothesis by fitting a linear model predicting the proportion of the population from ethnicity (centered, coded Not Hispanic = −0.5, Hispanic = 0.5), site (centered, coded Site 1 = −0.5, Site 2 = 0.5), and their interaction. The results of the regression indicate a significant effect of ethnicity [$b = -8.965$, $F(1, 78) = 903.474$, $p < 0.05$]. On average, there was a greater proportion of non-Hispanic respondents ($M = 0.930$) than Hispanic respondents ($M = 0.07$) across both sites. No other effects were significant (Table 4).

Educational Attainment

To assess differences in educational attainment between the two sites, we used the ACS 5-year estimates to separately compute the proportion of the population over 25 that has

received varying levels of education (High School Diploma equivalent or less, some college, Associate's degree, Bachelor's degree, or Graduate degree). The ACS estimates report educational attainment status for different age bands including 18–24 and 25+. We selected the estimates for the population over age 25 because the Facebook ad primarily reached an audience age between 25 and 44. For each respondent's census tract, we calculated five proportions (number of people within a respondent's census tract that attained an education level/total population within the census tract), corresponding to each of the education levels. We regressed the calculated proportions on educational attainment (dummy coded with Graduate degree as the reference group), site (centered, coded Site 1 = −0.5, Site 2 = 0.5), and their interaction. Site 2 represents a typical University-based convenience sample and therefore we expected

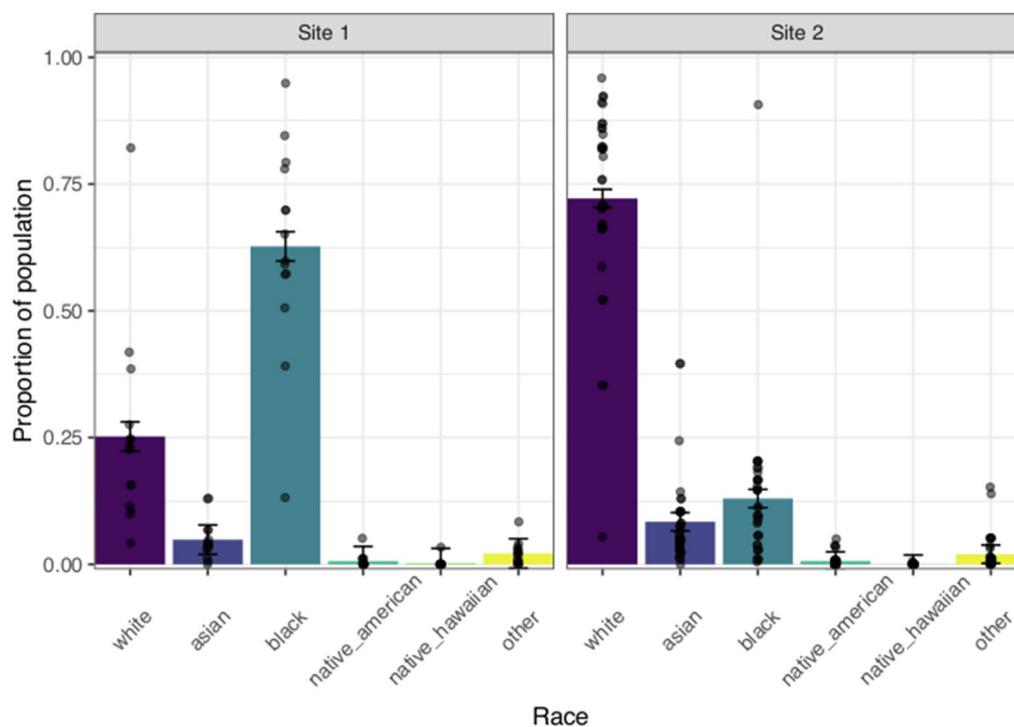


FIGURE 5 | Racial makeup by site location for respondents. Each bar represents the average proportion of each site's population that identifies as a particular racial category. The first panel shows mean values for Site 1 and the second panel shows mean values for Site 2. Each data point represents the proportion of the population that identifies as a particular racial category for individual respondents' census tracts. Error bars indicate SEM.

that respondents from Site 2 would have a higher proportion of highly educated people compared to Site 1.

The full results of the model analyzing educational attainment can be seen in **Table 4**. The variable educational attainment was contrast coded with Graduate degrees as the reference group, which resulted in four different comparisons. Each estimate indicates whether the proportion of the population that received a Graduate degree differs significantly from the proportion of the population that received each of the other levels of education (i.e., proportion of people that received a Graduate degree compared to the proportion of people that received a high school degree). Collapsing across sites, a higher proportion of individuals living in the respondents' census tracts reported having a High School diploma or less ($M=0.290$) or attended some college ($M=0.200$) compared to a Graduate degree ($M=0.180$). There was also a significant site by educational attainment interaction [$F(4, 240)=25.644$, $p<0.05$, $n^2p=0.299$]. As shown in **Figure 6**, Site 1 had a greater proportion of people that received a High School Diploma or less ($M=0.439$) than a Graduate degree ($M=0.081$) [$F(1, 240)=62.299$, $p<0.05$] compared to Site 2 ($M=0.227$ and $M=0.216$ for High School Diploma or less and Graduate degree, respectively). The same pattern of results can be seen in **Figure 6** for the proportion of people located in Site 1 that attended some college ($M=0.268$) rather than those who received a graduate degree ($M=0.081$) as compared to Site 2 ($M=0.172$ and $M=0.216$ for some college and Graduate degree, respectively). The analysis of educational attainment supports our prediction that targeting

a more diverse location can provide a sample of participants that have a wider range of educational backgrounds than typically seen in University-based samples.

Household Income

We identified the median household income estimates from the ACS for each census tract to provide a metric of socioeconomic diversity for respondents from the two sites. We predicted that Site 2 would have a higher median income than Site 1, reflecting the typical wealthy convenience sample. To investigate this hypothesis, we fit a linear model predicting median household income from site location. Site location significantly predicted median income [$b=23,428$, $F(1, 48)=14.13$, $p<0.05$, $n^2p=0.227$]. Site 2 had a higher median income ($M=68,997.33$) than Site 1 ($M=45,569.07$). These results indicate that targeting Site 1 resulted in a sample of respondents who are likely to be more economically diverse than would have been possible had we only targeted the local convenience sample.

Study 2 Discussion

Study 2 investigated whether targeting a Facebook advertisement to a more diverse location would provide a more representative pool of participants for future online research. We directed microtargeting Facebook advertisements toward two locations: a diverse urban community and a location proximal to our university that reflects typical local convenience sampling. The

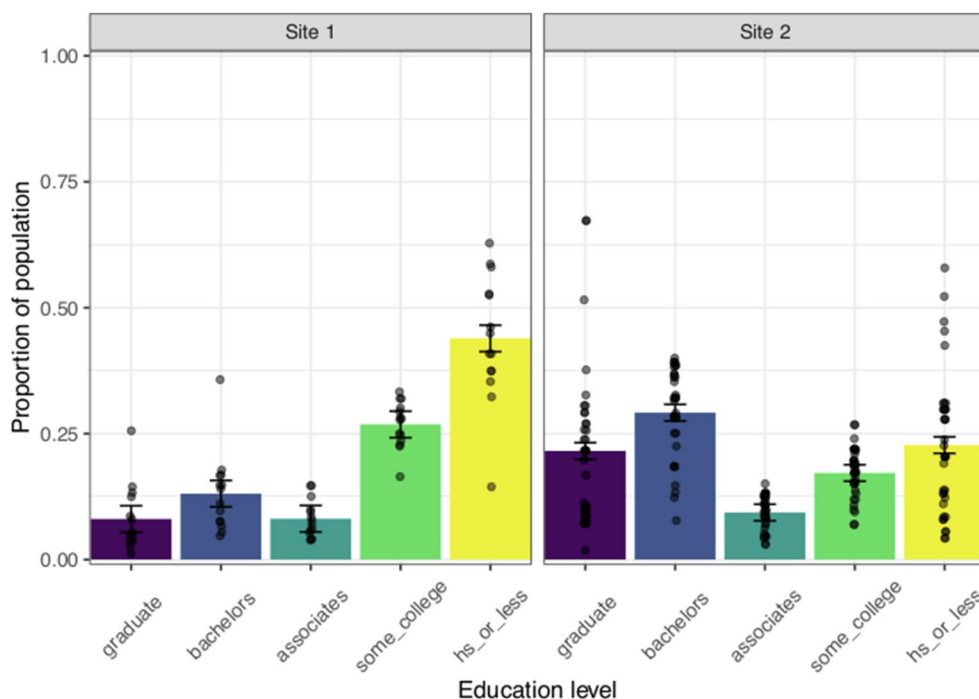


FIGURE 6 | Educational attainment by site location for respondents. Each bar represents the mean proportion of the population that has attained each education level for Site 1 (panel 1) and Site 2 (panel 2). Each data point represents the proportion of the population that has reached each education level for individual respondents' census tracts. Error bars represent the SEM.

advertisement had high engagement and provided 59 new sign-ups across the two site locations. Importantly, the analyses of our census tract-based diversity metrics suggest that the 14 respondents from Site 1 were likely to be more diverse in racial, educational, and economic backgrounds than the 37 respondents from Site 2. Our results lend credence to the potential benefits of recruiting representative samples for online studies using targeted Facebook ads. Further, these results suggest that widening the net of recruitment to more diverse locations can create a pool of participants for online studies who are more demographically representative than is possible for in-lab studies that are limited by the diversity of the local population. It remains to be seen, however, whether greater diversity in families who respond to our ads will lead to greater diversity in the families who eventually choose to participate in our studies.

GENERAL DISCUSSION

The present work demonstrates (1) the viability of the Zoom platform for experimenter-moderated looking time studies using LWL paradigms with infants, (2) the feasibility of online participation regardless of internet speed, and (3) the effectiveness of microtargeted Facebook ads for recruiting a more diverse group of potential participants. Overall, the current research demonstrates not only just the feasibility of running studies with infants online with this paradigm, but also addresses

some of the immediate concerns surrounding recruitment diversity and data quality.

Caregivers were able to appropriately set-up their computer for the study with the virtual aid of the experimenter and deploy the experiment themselves while the experimenter recorded and stored the participant data. This method did not sacrifice data quality and was easy to administer. Although access to high-speed internet was a paramount concern prior to online data collection, the current study suggests that internet connectivity does not significantly reduce data quality. Lower speed internet can impact the experiment presentation, but the Zoom recording captures these perturbations. For example, the experiment did not display the first trial when running the study using 2G internet. The experimenter-facing Zoom recording reflected this presentation error and looking behavior was not coded during the missed trials. We also anticipated that internet speed would significantly impact the validity of the timing of some participants' data. However, we were able to accommodate this variability by individually adjusting the frame rate for each trial prior to data analysis. Together, these findings demonstrate that online data collection can yield similar results to in-lab studies without significant restrictions due to participant internet connectivity.

Virtual study administration is accompanied by concerns regarding equity in internet access. For optimal study administration and for the clearest data quality, faster internet speeds are optimal; however, this does not mean that slower internet speeds preclude participation. Participants in the

sample had varied internet speeds, but that did not prohibit them from participation. As demonstrated by our internet speed testing results, there is minimal data loss at even the slowest internet speed, and the data loss that is incurred is present on both the participant and experimenter sides.

One disadvantage to online research using Zoom is that people cannot participate on tablets and smartphones, whereas TheChildLab can be used on these devices (Sheskin and Keil, 2018). Experimenter-moderated Zoom-based eye-gaze tasks like Study 1 require a desktop computer or laptop with a web camera and internet access to participate. The screen sharing function on Zoom does not allow for simultaneous screensharing and video sharing on tablets or smartphones. These constraints will prevent a segment of the population from having access to participating in research like Study 1. According to the National Center for Education Statistics (2020), 6% of 3- to 18-year-old only have home access to the internet *via* smartphone, with an additional 6% of children having no internet access at home. Most of the children without access to the internet *via* a device other than a smartphone are from minority groups, have parents with the equivalent of a high school diploma or less, and are from the lowest quarter of all family incomes (National Center for Education Statistics, 2020). In other countries, lack of access may be substantially greater. Lourenco and Tasimi (2020) suggest several ways to combat these limits on research participation, including mobile laboratory set-ups to go into communities with less internet access and providing mobile hotspots to participant families for participation. These approaches may facilitate recruiting representative participant samples, as 12% of the child participant population is currently unreachable *via* the Zoom videoconferencing online methodology.

Online recruitment is not enough to check the diversity box, as is evident in the highly non-representative sample in Study 1. Online recruitment efforts require intentionality in making decisions on the locations to target and the materials included in the ads. Microtargeted Facebook (and Instagram) ads work for caregiver, and subsequent infant, recruitment. The results of Study 2 suggest that we may have reached more diverse respondents *via* recruitment efforts focused on specific area codes. However, because we have not yet enrolled these respondents in studies, additional research is needed to verify that these recruitment efforts subsequently result in more representative study participants. Further, the microtargeted Facebook ads used in the current study depicted a White infant with a multiracial researcher and a Black infant with both a Black and a White researcher. These advertisement design decisions may have increased the level of response by non-White caregivers, perhaps because they saw people that look like themselves and their child(ren) represented in a research setting. Indeed, findings from the marketing literature demonstrate positive relationships between the amount of diversity presented in recruitment materials and recruitment of more diverse job candidates (Avery et al., 2004; Walker et al., 2012).

Limitations and Future Directions

Study 1 demonstrates that experimenter-moderated LWL tasks are feasible *via* the Zoom platform. However, the conclusions that we can draw about the timing of fixations is limited by the comparisons we can draw. Because we do not have an identical in-lab task to which we can compare the timing data, we compared our data to other in-lab LWL studies reported on Peekbank (Zettersten et al., 2021). While this is a helpful comparison, many features diverge between our task and these extant data (e.g., number of trials testing each word). Additionally, the set-ups of virtual and in-lab studies differ tremendously in the positioning of the child relative to the screen, as well as the size of the screen on which the study is administered. If administered in our current lab set-up, this study would have been presented on a 55-in Toshiba LCD television with participants seated on their caregiver's lap 3 feet away from the screen. In the virtual experimenter-moderated version reported in Study 1, the task was administered on a 13- to 15-in computer screen with the participants approximately 1 foot away from the screen. In both environments, objects on the screen are evenly spaced on the left and right of the screen, but the size of the objects and the distance between them differs as a function of the size of the monitor. This may account for some of the looking time differences between Study 1 and the Peekbank comparison – the distance between objects impacting the amount of time it takes to complete a saccade.

In Study 2, recruitment efforts *via* selected diverse photos and microtargeting diverse zip codes led to respondents from more diverse locations, but this does not necessarily beget diverse study participants. In line with what we had predicted, there was more diversity in the respondents from Site 1, though the overall number of respondents from Site 1 was less than half than the number of respondents from Site 2. This aligned with our concern of whether people from a more diverse area, and an area that is non-local to the University, would be willing to sign up to participate in an online study due to historical mistrust of research. In the future, additional ad specificity would allow a better understanding of the degree to which microtargeted recruitment increases the diversity of participant samples. This would also supply added insight into the remaining barriers for diverse participation. Microtargeted ads using the Facebook ad platform are accessible from mobile devices and tablets, though a mobile-device is not compatible with the present experimenter-moderated study administration. The requirement of a computer with a web camera and internet access places an added burden on the participants' families and may be a hindrance to study participation, despite sign-up interest.

In sum, conducting studies online provides a wider range of participant families the opportunity to partake in research, without researchers sacrificing data quality due to internet connectivity. The Zoom videoconferencing platform is widely available to caregivers and provides an easy avenue for experimenter-moderated eye-gaze studies using LWL. Moving

forward with online data collection requires intentionality on the part of the researchers to ensure they are recruiting diverse participants by using thoughtfully constructed recruitment materials, including the photos and language used. These efforts, combined, allow data collection to continue at a distance, and move us closer to samples that are more representative of the demographics of the population. The present work demonstrates not only the success, but also the feasibility of these efforts.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/e9a8y/?view_only=f74b7fbdcf04cdda13b479cf08a4c67

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The University of Wisconsin-Madison Educational and Social Behavioral Science Institutional Review Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin, and obtained from the individual(s) and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

REFERENCES

- Adolph, K. E., Gilmore, R. O., and Kennedy, J. L. (2017). Video data and documentation will improve psychological science. *Psychol. Sci. Agenda* 31.
- American Psychological Association (2003). Guidelines on multicultural education, training, research, practice, and organizational change for psychologists. *Am. Psychol.* 58, 377–402. doi: 10.1037/0003-066X.58.5.377
- American Psychological Association, APA Task Force on Race and Ethnicity Guidelines in Psychology (2019). Race and ethnicity guidelines in psychology: promoting responsiveness and equity. Available at: <https://www.apa.org/about/policy/guidelines-race-ethnicity.pdf> (Accessed June 15, 2021).
- Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., and Lawless, M. (2019). Using zoom videoconferencing for qualitative data collection: perceptions and experiences of researchers and participants. *Int. J. Qual. Methods* 18:1609406919874596. doi: 10.1177/1609406919874596
- Avery, D., Hernandez, M., and Hebl, M. (2004). Who's watching the race? Racial salience in recruitment advertising. *J. Appl. Soc. Psychol.* 34, 146–161. doi: 10.1111/j.1559-1816.2004.tb02541.x
- Bergelson, E., and Aslin, R. N. (2017). Nature and origins of the lexicon in 6-month-olds. *Proc. Natl. Acad. Sci. U. S. A.* 114, 12916–12921. doi: 10.1073/pnas.1712966114
- Bergelson, E., and Swingle, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proc. Natl. Acad. Sci. U. S. A.* 109, 3253–3258. doi: 10.1073/pnas.1113380109
- Brodeur, M. B., Guérard, K., and Bours, M. (2014). Bank of standardized stimuli (BOSS) phase II: 930 new normative photos. *PLoS One* 9:e106953. doi: 10.1371/journal.pone.0106953
- Buhrmester, M., Kwang, T., and Gosling, S. D. (2016). "Amazon's mechanical Turk: a new source of inexpensive, yet high-quality data?" in *Methodological*

AUTHOR CONTRIBUTIONS

DB and HW designed the study with critical insight from JS and wrote the manuscript. HW performed the research and analyzed the data. DB, HW, and JS contributed to revisions of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1747503 to DB. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work was also supported by grants from NICHD awarded to JS (R37HD037466) and the Waisman Center (U54 HD090256).

ACKNOWLEDGMENTS

We would like to thank the participating families as well as members of the Infant Learning Lab. In particular, we thank Ron Pomper for insight into Peyecoder, Martin Zettersten for his insight into Peekbank, Emily Kassens for her help with data collection, and Adlina Mohamed Muhrizan and Ellie McIntosh for their help with data coding. We thank Ellie Breitfeld for comments on a previous draft.

- Issues and Strategies in Clinical Research*. ed. A. E. Kazdin (Washington, DC: American Psychological Association), 133–139. doi: 10.1037/14805-009
- Casler, K., Bickel, L., and Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput. Hum. Behav.* 29, 2156–2160. doi: 10.1016/j.chb.2013.05.009
- Curtin, J. (2018). lmsupport: support for linear models. *R package version 2.9.13*. Available at: <https://CRAN.R-project.org/package=lmsupport> (Accessed August 4, 2021).
- Dworkin, J., Hessel, H., Gliske, K., and Rudi, J. (2016). A comparison of three online recruitment strategies for engaging parents: online recruitment. *Fam. Relat.* 65, 550–561. doi: 10.1111/fare.12206
- Enders, C. K. (2013). Dealing with missing data in developmental research. *Child Dev. Perspect.* 7, 27–31. doi: 10.1111/cdep.12008
- Fernald, A., Pinto, J. P., Swingle, D., Weinberg, A., and McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychol. Sci.* 9, 228–231. doi: 10.1111/1467-9280.00044
- Fernald, A., Zangl, R., Portillo, A. L., and Marchman, V. A. (2008). "Looking while listening: using eye movements to monitor spoken language comprehension by infants and young children," in *Language Acquisition and Language Disorders: Vol. 44. Developmental Psycholinguistics: On-Line Methods in Children's Language Processing*. eds. I. A. Sekerina, E. M. Fernández and H. Clahsen (Philadelphia, PA: John Benjamins Publishing Company), 97–135. doi: 10.1075/lald.44.06fer
- Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., and Yurovsky, D. (2016). Using tablets to collect data from young children. *J. Cogn. Dev.* 17, 1–17. doi: 10.1080/15248372.2015.1061528
- Garrison, H., Baudet, G., Breitfeld, E., Aberman, A., and Bergelson, E. (2020). Familiarity plays a small role in noun comprehension at 12–18 months. *Infancy* 25, 458–477. doi: 10.1111/infa.12333

- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not WEIRD. *Nature* 466:29. doi: 10.1038/466029a
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.
- Jang, M., and Vorderstrasse, A. (2019). Socioeconomic status and racial or ethnic differences in participation: web-based survey. *JMIR Res. Protoc.* 8:e11865. doi: 10.2196/11865
- Klein-Radukic, S., and Zmyj, N. (2015). Dropout in looking time studies: The role of infants' temperament and cognitive developmental status. *Infant Behav. Dev.* 41, 142–153. doi: 10.1016/j.infbeh.2015.10.001
- Krogstad, J. M. and Pew Research Center (2015). Social media preferences vary by race and ethnicity. Available at: <https://www.pewresearch.org/fact-tank/2015/02/03/social-media-preferences-vary-by-race-and-ethnicity/> (Accessed March 7, 2021).
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Mahr, T., McMillan, B. T. M., Saffran, J. R., Ellis Weismer, S., and Edwards, J. (2015). Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition* 142, 345–350. doi: 10.1016/j.cognition.2015.05.009
- National Center for Education Statistics (2020). The condition of education – preprimary, elementary, and secondary education – family characteristics – children's internet access at home – indicator May (2020). Available at: https://nces.ed.gov/programs/coe/indicator_cch.asp#info
- National Institutes of Health (1994). NIH guidelines on the inclusion of women and minorities as subjects in clinical research. *Fed. Regist.* 59, 14508–14513.
- Nicholson, J., Deboeck, P., and Howard, W. (2015). Attrition in developmental psychology: a review of modern missing data reporting and practices. *Int. J. Behav. Dev.* 41, 143–153. doi:10.1177/0165025415618275
- Oakes, L. M. (2010). Editorial comment: infancy guidelines for publishing eye-tracking data. *Infancy* 15, 1–5. doi: 10.1111/j.1532-7078.2010.00030.x
- Olson, R. H., Pomper, R., Potter, C. E., Hay, J. F., Saffran, J. R., Ellis Weismer, S., et al. (2020). Peyecoder: an open-source program for coding eye movements (version v1.1.5). *Zenodo*. doi: 10.5281/zenodo.3939234
- Pechmann, C., Phillips, C., Calder, D., and Prochaska, J. J. (2020). Facebook recruitment using zip codes to improve diversity in health research: longitudinal observational study. *J. Med. Internet Res.* 22:e17554. doi: 10.2196/17554
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/> (Accessed August 4, 2021).
- Scott, K., and Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Sheskin, M., and Keil, F. (2018). TheChildLab.com a video chat platform for developmental research. *PsyArXiv* [Preprint]. doi:10.31234/osf.io/rn7w5
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Smith-Flores, A. S., Perez, J., Zhang, M. H., and Feigenson, L. (2021). Online measures of looking and learning in infancy. *PsyArXiv* [Preprint]. doi:10.31234/osf.io/tdbnh
- Swingle, D., and Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition* 76, 147–166. doi: 10.1016/S0010-0277(00)00081-0
- Swingle, D., and Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychol. Sci.* 13, 480–484. doi: 10.1111/1467-9280.00485
- United States Census Bureau (2020). American Community Survey 5-year data (2009–2019). Available at: <https://www.census.gov/data/developers/data-sets/acs-5year.html> (Accessed February 2021).
- Venker, C. E., Pomper, R., Mahr, T., Edwards, J., Saffran, J., and Ellis Weismer, S. (2020). Comparing automatic eye tracking and manual gaze coding methods in young children with autism spectrum disorder. *Autism Res.* 13, 271–283. doi: 10.1002/aur.2225
- Walker, H., Feild, H., Bernerth, J., and Becton, J. (2012). Diversity cues on recruitment websites: investigating the effects on job seekers' information processing. *J. Appl. Psychol.* 97, 214–224. doi: 10.1037/a0025847
- Wass, S. V., Smith, T. J., and Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behav. Res. Methods* 45, 229–250. doi: 10.3758/s13428-012-0245-6
- Yurovsky, D., and Frank, M. C. (2017). Beyond naive cue combination: salience and social cues in early word learning. *Dev. Sci.* 20:e12349. doi: 10.1111/desc.12349
- Yurovsky, D., Wade, A., and Frank, M. C. (2013). "Online processing of speech and social information in early word learning," in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, July 31–Aug 3, 2013. eds. Knauff, M., Pauen, M., Sebanz, N., and Wachsmuth, I. (Berlin, Germany), 1641–1646.
- Zettersten, M., Bergey, C., Bhatt, N. S., Boyce, V., Braginsky, M., Carstensen, A., Frank, M. C. (2021, May 12). Peekbank: Exploring children's word recognition through an open, large-scale repository for developmental eye-tracking data. doi: 10.31234/osf.io/ep693s

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Bacon, Weaver and Saffran. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Agreement and Reliability of Parental Reports and Direct Screening of Developmental Outcomes in Toddlers at Risk

Juan Giraldo-Huertas^{1*} and Graham Schafer²

¹ Department of Psychology of Development and Education, Universidad de la Sabana, Chía, Colombia, ² The School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

OPEN ACCESS

Edited by:

Rhodri Cusack,
Trinity College Institute
of Neuroscience, Ireland

Reviewed by:

Alessandra Geraci,
University of Trento, Italy
Angela Conejero,
University of Granada, Spain

*Correspondence:

Juan Giraldo-Huertas
juangh@unisabana.edu.co

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 15 June 2021

Accepted: 03 September 2021

Published: 28 September 2021

Citation:

Giraldo-Huertas J and Schafer G
(2021) Agreement and Reliability
of Parental Reports and Direct
Screening of Developmental
Outcomes in Toddlers at Risk.
Front. Psychol. 12:725146.
doi: 10.3389/fpsyg.2021.725146

Developmental screening is a practice that directly benefits vulnerable and low-income families and children when it is regular and frequently applied. A developmental screening tool administered by parents called CARE is tested. CARE contains a compilation of activities to report and enhance development at home. Hundred and fifty-seven families in Bogotá (Colombia) initially responded to a call to participate in developmental screening tools' validation and reliability study. All children (Average: 42.7 months old; *SD*: 9.4; Min: 24, Max: 58) were screened directly by trained applicants using a Spanish version of the Denver Developmental Screening test [i.e., the Haizea-Llevant (HLL) screening table]. After a first screening, 61 dyads were positive for follow-up and received a second HLL screening. Fifty-two out of 61 dyads use and returned CARE booklet after 1-month screening at home. The comparative analysis for parent reports using CARE and direct screening observation included (a) the effects of demographic variables on overall and agreement, (b) agreement and congruence between the CARE report classification and direct screening classification ("At risk" or "Not at risk"), (c) receiver operating characteristic analysis, (d) item-Level agreement for specific developmental domains, and (e) acceptability and feasibility analysis. Results and conclusions show the parental report using the CARE booklet as a reliable screening tool that has the potential to activate alerts for an early cognitive delay that reassure clinicians and families to further specialized and controlled developmental evaluations and act as a screen for the presence of such delay in four developmental dimensions.

Keywords: parental reports, developmental screening, children at risk, reliability and agreement studies, low-middle income countries, receiver operating characteristic (ROC) analysis

INTRODUCTION

Attention to screening tools in low-and-middle income countries (ongoing: LMIC) settings has grown recently (Boggs et al., 2019). However, only population-level tools (i.e., instruments for monitoring countries or regional status) have been shown to have acceptable accuracy, reliability, and feasibility for routine use in health and educational systems. Individual-level tools (i.e., instruments to measure cases or single participant assessment) are not frequently reported to have utility in planning for direct early interventions. Efforts for optimal monitoring and screening tools

have a direct relationship with the Nurturing Care Framework (Britto et al., 2017; WHO, 2020). The Nurturing Care Framework has inspired a considerable literature for early interventions in LMIC (Trude et al., 2021). Reviews of previous screening and surveillance projects around parenting effects on children development, shown how high nurturing interventions reduce negative effects of scarce and adverse environments (Lu et al., 2020; Tann et al., 2021). However, there is no complete or permanent program in an LMIC that ensures constant and relevant evidence-based approaches to monitoring and assessment of child development or nurturing status (Milner et al., 2019). Along with monitoring, even in high income countries, indicators and information to design interventions and programs guided by developmental screening (DS) to reduce social and educational inequity are incomplete (NASEM, 2019). The NASEM report showed how, before the COVID-19 pandemic, standard health information systems needed improvements in research and data sources, to fill important gaps in knowledge about child intervention programs to identify promising program features to implement effectively at scale. The same efforts are needed in getting accurate information including a call for action through developmental monitoring and screening in LMIC (Goldfeld and Yousafzai, 2018). Increasing developmental monitoring and screening of children's outcomes can optimize early intervention referrals, assessments, and eligibility (Barger et al., 2018). Also, in LMIC like Colombia, where this pilot study take place, screening tools for children monitoring about developmental risks should fight against the impact of social inequalities in children's development, a primary socio-political goal and where testing children directly by public administration services it is not always accessible in vulnerable populations (Rubio-Codina and Grantham-McGregor, 2020).

The main aim in the present study is related to the Compilation of Activities to Report and Enhance development (Ongoing: CARE), a booklet created to obtain screening information of daily activities of interaction between parents or caregivers with children in vulnerable families living in Colombia. The consequent aims of the current study are threefold:

- (1) Explore the diagnostic characteristics and performance of CARE as a tool for DS using parent reports, with item agreement analysis at the individual level between parent reports and direct assessment in particular domains, as set out above.
- (2) Examine consistency between parental reports using CARE and classification and scores using an external screening in the domains of personal-social skills, language and logico-mathematical reasoning, fine motor-adaptive and gross motor skills. We expect to find similar results to prior research showing good agreement between parent report and direct testing of social, language and gross motor skills, but somewhat weaker agreement in fine motor skills (Miller et al., 2017).
- (3) Obtain relevant data to identify the validity of CARE, with feedback of the findings to both academic and institutional administrators engaged in participant enrollment.

Following paragraphs extend the rationale for every specific aim.

The first aim explores the diagnostic characteristics and performance of a new DS tool administrated by parents and compared with an external screening tool measurement. Improving screening and developmental status measurement in early child development is feasible, but several coverage and quality characteristics remain unreachable for evidence-based interventions in LMIC (Milner et al., 2019). Interventions with simpler, routinary and including multi-domain outcome measurement needs well-designed tools. DS tools reduce financial and time costs for fundamental research and public health activities, such as assessing early developmental status at an individual level (Johnson et al., 2008), even in LMIC (Tann et al., 2021). However, several decision-making steps are required when DS tools are included in interventions, monitoring programs, assessments, or research (Nadeem et al., 2016). In the last decade, different studies have evaluated DS tools deployed at primary healthcare services in LMIC (Fischer et al., 2014; Fernald et al., 2017; Boggs et al., 2019). These three studies rated 14 individual-level tests, applying common criteria for validity, reliability, accessibility of application, required training, administration time, cultural adaptability, geographical uptake, and clinical relevance and utility. Utility was only considered for the category of individual-level measurement tools. Boggs et al. (2019) excluded the costs of the tool (i.e., the budget necessary to buy and use the materials and to train personnel) from the criteria listed by Fischer et al. (2014). The review of 14 individual-level tests indicated higher ratings of administration time or reliability compared with population-level and ability-level tools (Boggs et al., 2019). Of these 14 individual-level tests, 36% ($n = 5$) had a higher rating for both administration time and reliability: namely, the Ages and Stages Questionnaire (ASQ), the Denver Developmental Screening Test (DDST), the Guide for Monitoring Child Development (GMCD), the ICMR Psychosocial Development Screening Test, and the Parents' Evaluation of Developmental Status (PEDS). Those review studies did not find any screening tool that was particularly used or designed in Colombia (Fischer et al., 2014; Boggs et al., 2019).

The Colombia's Ministry of Health uses the Abbreviated Development Scale (Ongoing ADS; in Spanish, *Escala Abreviada del Desarrollo*; Ortiz, 1991) not like a screening tool, but in different institutional scenarios, including children's centers and public kindergartens around the country, to obtain information about children emotional, cognitive and health conditions. Colombia's Ministry of Health (Ministerio de Salud de la República de Colombia, 2016) presents the ADS with no published report on its conceptualization, pilot testing, or complete analysis of validity and reliability. A partial validation analysis of the ADS-1 for the language and hearing domain in 4- to 5-year-old children indicated low predictive ability (Sensitivity: 54%, Specificity: 42%) and poor agreement with a gold standard for early detection of language and hearing disorders (i.e., the *Reynell norm-referenced test*) on measuring expressive and receptive language skills, and with tone audiometry and otoacoustic emissions on assessing hearing (Muñoz Caicedo et al., 2013). We can therefore conclude that to the best

of our knowledge, it is not a well-designed tool for the Colombian context, following the standards of Boggs et al. (2019). Moreover, the aforementioned rating exercises report the use of a “developmental domain” approach to the relevant screening tools, but not an analysis of “administration of test,” which is recommended by different authors (Fernald et al., 2017; Boggs et al., 2019). The “administration of test” view implies comparing caregiver reports with direct child observation. Vitrikas et al. (2017) described both a parent-completed DS tool as an instrument for obtaining screening information through parent participation, and (as a separate instrument) a directly administered DS tool when information is based on direct observation of the child by a physician or other expert.

The second of the three aims examine the consistency between parental reports using CARE and classification and scores using an external administrated tool, including reliability and agreement analysis. DS still has some unique challenges associated with obtaining accurate data in early childhood, especially in LMIC and families in poverty conditions (Lu et al., 2020). The Early Childhood Development Index (ECDI), for example, is a 10-question survey used in the Nurturing Care Framework to determine whether children are on track in their cognitive and social-emotional development (Richter et al., 2017, 2020). For global, national, and regional level, ECDI information is fundamental, but high-quality and comparable data for individual developmental status is not fully captured by developmental surveys or questionnaires (McCoy et al., 2016, 2018; Lu et al., 2020). Parental reports are a high-quality, reliable alternative to obtaining individual child information via home visits. We define ‘parent report’ in this study as information obtained from a parent using CARE®. The CARE is a booklet created to obtain information of daily activities of interaction between parents or caregivers with children, derived from an instrument applied by training specialized personal in a 3-year research program, with a sample of 1173 children under 6 years old and their caregivers in two large territorial regions of Colombia (Cundinamarca and Boyacá), in urban and rural settings (Giraldo-Huertas et al., 2017). The main content of CARE includes activities to report developmental milestones in four domains mentioned before, for two age groups: 24–35 months old and 36–59 months old. Every item in CARE is closely related to one item in the Haizea-Llevant (HLL) Table (Iceta and Yoldi, 2002). The HLL screening table is a DS tool derived from the Denver Developmental Screening Test (DDST) and the Denver Pre-screening Developmental Questionnaire (Frankenburg et al., 1976; Frankenburg, 1987). HLL was selected because the DDST is broadly used and standardized in different countries (Lipkin and Gwynn, 2007; Guevara et al., 2013; Dawson and Camp, 2014), including populated regions in Brazil (Lopez-Boo et al., 2020) and Colombia (Rubio-Codina and Grantham-McGregor, 2020). The HLL is a similar Spanish language version of the DDST, used previously in a long-term health screening program in the Basque Country (Fuentes-Biggi et al., 1992; Rivas et al., 2010). The HLL items included in CARE and the whole designing process follow the components recommended by Nadeem et al. (2016) for construction and validation of assessment tools. Conceptualization and consolidation phases

were realized in the IPV (Inicio Parejo de la Vida, “Equal Start in Life”), a research program with previously take place in Colombia (Giraldo-Huertas et al., 2017).

Compare parental reporting and direct assessment are defined as the two main methods used to evaluate child development (Miller et al., 2017). Miller et al. (2017) remark on the need to determine reliability and agreement in parental reports in the early detection of developmental delays, comparing these with direct assessments as a quality control procedure. In a framework for optimal quality in early childhood assessments, reliability and agreement (R&A) studies are often expected (Vanbelle, 2017). R&A studies provide information about the quality of measurements, specifically about the ability of a scale to differentiate between the items, despite the presence of measurement error (reliability); and also, about the degree of closeness between two assessments made on the same items (agreement). Good levels of R&A are essential for new measurement tools if they are to be included in clinical decision making and subsequent interventions (Vanbelle, 2017). R&A application may relieve technical concerns about the accuracy of parental reporting (Bennetts et al., 2016; Miller et al., 2017). Parents are an important source of information regarding child skill deficits and atypical behaviors, because they are uniquely positioned to observe and interact with children across various daily interactions at home (Jeong et al., 2019). Also, for developmental monitoring (i.e., healthcare professionals’ practices to make informed clinical judgments about children’s developmental progress based on their own criteria) parent reports might be included to help identify children at risk (Barger et al., 2018; Gellasch, 2019). Developmental monitoring practices with parent reports for individual developmental status and later diagnostic testing may be shorter to administer, thereby reducing costs and increasing developmental delay identification in the regular health visits at 9, 18, and 24–30 months (Miller et al., 2017; Vitrikas et al., 2017; Gellasch, 2019).

Finally, a third aim is to obtain relevant data to identify the validity of CARE in protocols for feedback of the use and individual results to both academic and institutional administrators engaged in participant enrollment. Unfortunately, even in high-income countries, only a small proportion of children regularly receive developmental monitoring in health systems, preventing the detection of early delays and subsequent interventions (Barger et al., 2018). The COVID-19 pandemic may have exacerbated adversity and imposed still more barriers to the optimization of developmental monitoring (Richter et al., 2020; Trude et al., 2021), making parental reports valuable tools for identifying individual children’s developmental status. The present study aims to evaluate consistency between two sources of information—direct assessment and parent report—when classifying at-risk children and measuring child development in four domains (personal-social, language and logico-mathematical reasoning, fine motor-adaptive, and gross motor skills) within a reliability and agreement analysis, and finally, a validity report for inclusion in future institutional or community scenarios.

It is important to note that the parental administration method does not profess to replace any clinical or scientific

intervention and will presumably run in parallel with other previously existing or subsequently developed screening and intervention methods for health and educational systems. Specifically, this study review CARE characteristics and initial scopes as a screening tool, and it is not possible to currently consider that should be used for intervention.

MATERIALS AND METHODS

Participants

Participants were dyads of toddlers and principal caregivers recruited at a children's center pertaining to a community-level social support intervention that was part of a wider government-funded nutritional program. The study's catchment area included an urban population vulnerable to poverty in the north-west of Bogotá, Colombia. One hundred and fifty-seven families ($N = 157$) initially responded to a call to participate in a study of tools for a future cognitive intervention and completed documentation for informed consent (Figure 1). All children were screened using the HLL screening table (Iceta and Yoldi, 2002). Due to reported application practices for early DS (Alcantud et al., 2015), HLL was applied twice. The first application intent to diminish possible anxiety or fear around working with a health professional in screening settings (Villagomez et al., 2019) and follows the recommended application twice before screening decisions with participants in systems for early detection of developmental disorders (Alcantud et al., 2015). One week later after a first screening with HLL, 61 dyads (85.2%) were positive for follow-up and received a second HLL screening. Some 52 caregivers out of these 61 dyads returned the CARE booklet after using it as a screening tool at home.

The sample included all families who satisfied the following criteria: (1) They had at least one pre-school child (aged 59 months or younger); (2) they were currently in a couple, unless it was unfeasible to talk with one partner (excluding, e.g., partners who traveled a lot, widows, divorcees; (3) they understood written or spoken Spanish; and (4) they were willing

to receive a CARE booklet and use it as a screening tool, to the best of their capabilities. Sociodemographic characteristics of the final participants sample are described in Table 1. The procedure to obtain sociodemographic information, described below, does not establish any statistical difference in the profile of families who dropped out of the study at different stages.

Measures

Each dyad was interviewed and received:

- (1) Sociodemographic information survey (The Questionnaire for Parents and Caregivers General Data; Profamilia, 2010; Giraldo-Huertas et al., 2017).
- (2) The Haizea-Llevant screening table (Iceta and Yoldi, 2002).
- (3) The CARE booklet.

The Questionnaire for Parents and Caregivers General Data

The Questionnaire for Parents and Caregivers General Data (GDQ) was used in the IPV (Inicio Parejo de la Vida—Equal Start in Life) program (Giraldo-Huertas et al., 2017) and contains the 14 variables associated with the socio-cognitive development of children of under 6 years of age in the geographic region of interest, including items from the ENDS (Encuesta Nacional de Demografía y Salud—Colombian National Survey of Demographics and Health; Profamilia, 2010). The GDQ comprises 68 questions in eight modules that obtain data about the social, demographic and health characteristics of children under 6 years-old and their families. All questions were answered by the mother or primary caregiver of each child. The survey took approximately half an hour per participant.

The Haizea-Llevant Screening Table

The HLL (Fuentes-Biggi et al., 1992; Iceta and Yoldi, 2002; Rivas et al., 2010) was used by the research team for individual assessment of children. The individual developmental performance score is defined as the number of age-appropriate test items of a domain in HLL that a child can successfully

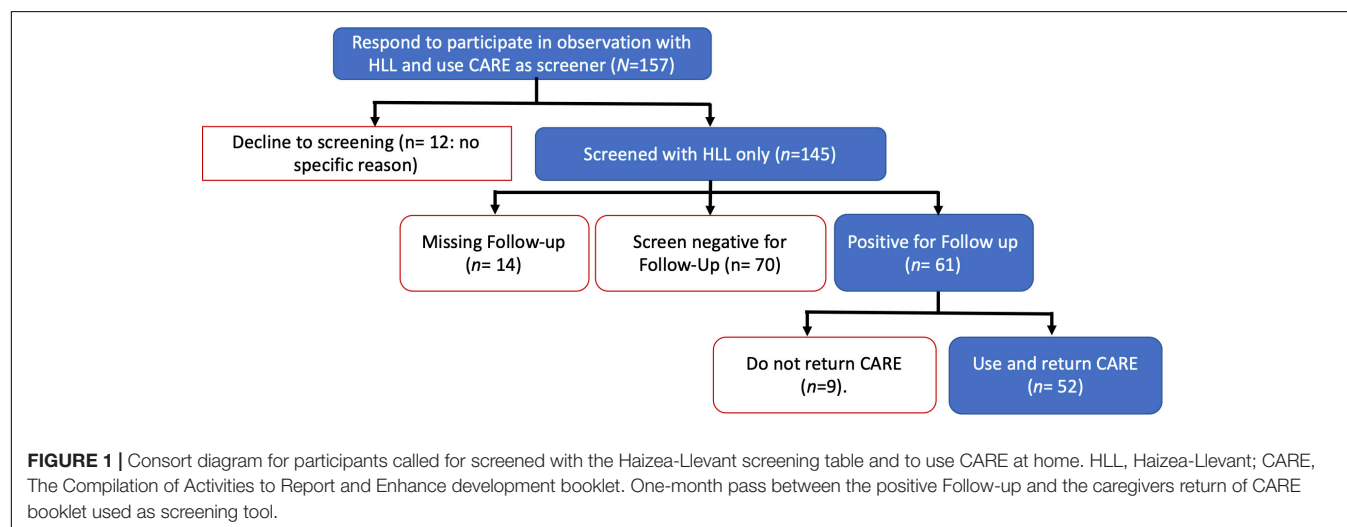


TABLE 1 | Characteristics of the sample for validation of CARE® (*n* = 52).

| Sex of the child | <i>n</i> (%) |
|---|---------------------|
| Female | 23 (44.2) |
| Male | 29 (55.8) |
| Age group | |
| 24–35 months old | 9 (17.3) |
| 36–47 months old | 25 (48.1) |
| 48–59 months old | 18 (34.6) |
| Principal caregiver (PC) | |
| Mother | 29 (55.8) |
| Relative at home | 9 (17.3) |
| Relative out of home | 5 (9.6) |
| Non-relative at home | 2 (3.8) |
| Non-relative out of home | 1 (1.9) |
| No answer | 6 (11.5) |
| PC educational level | |
| No school experience | 1 (1.9) |
| Incomplete elementary | 6 (11.5) |
| Elementary | 5 (9.6) |
| Incomplete high school | 2 (3.8) |
| High school | 18 (34.6) |
| Technician | 9 (17.3) |
| Incomplete undergraduate | 1 (1.9) |
| Undergraduate | 3 (5.8) |
| Postgraduate | 1 (1.9) |
| No answer | 6 (11.5) |
| Maternal Employment | |
| Employed | 34 (65.4) |
| Unemployed | 12 (23.1) |
| No answer | 6 (11.5) |
| Type of settlement | |
| Urban | 39 (75.0) |
| Non-urban | 4 (7.7) |
| No answer | 9 (17.3) |
| Socioeconomic national scale⁺ | |
| Level 1 Very low: Between 1488 and 1606 US Dollar by year or less. | 13 (25.0) |
| Level 2 Low: More than 1606 US Dollar by year but less than one national minimum wage (3.751 USD per year). | 19 (36.5) |
| Level 3 Medium low ⁺⁺ : less or more than one or two national minimum wage as household income. | 14 (27.0) |
| No answer | 6 (11.5) |

⁺Income are exchanged to US dollars in July/2020; ⁺⁺ Sources: MESEP-DNP (2011) and Sánchez-Torres (2015).

pass or not. For nominal classification, a “Caution” is recorded when an age-appropriate item is not passed. If the child is older than the limit age for the 95% of the standardization population passing the item, and does not pass it, that item is recorded as a “Delay.” As example for an item (“Identify colors”) in the domain of language and logic-mathematical reasoning: if a child is 40 months old and does not identify colors when these are pointed out by the interviewer, this is interpreted as a “Caution” item (**Figure 2A**); if a child is over 44 months old and does

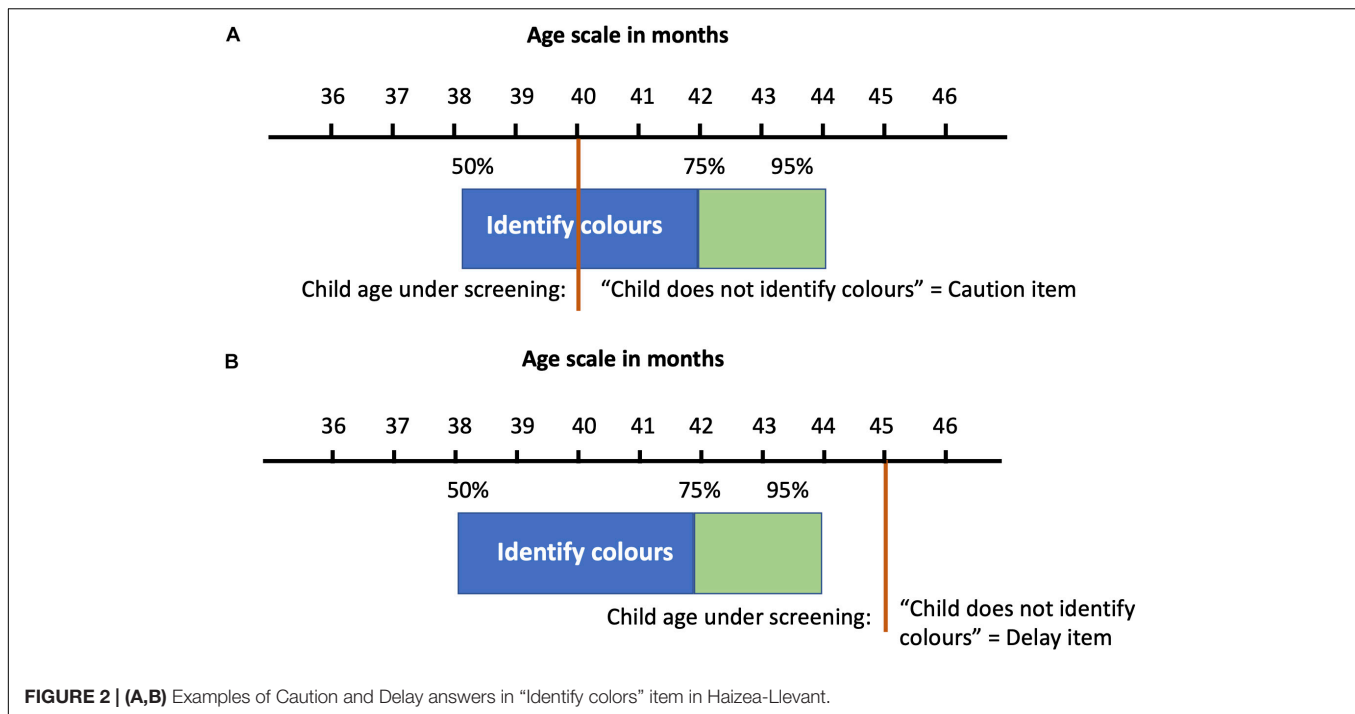
not identify colors during the observation with the HLL, this is interpreted as “Delay” item (**Figure 2B**).

The counting of Caution and Delay items enables scoring of the overall test and helps the interpretation of the screening, permitting additional evaluations and referrals as appropriate (Vitrikas et al., 2017). For nominal classification of the results, if the child at least one Delay item or at least two Cautions, he/she would be classified “At risk.” No Delay answers and just one Caution answer would lead to a classification of “Passing.” Henceforth, we classify those participants “Passing” the HLL as “Not at risk.” For developmental domain analysis, values were scored following a recent approach for the Denver II test, using an analysis of the distribution of items in the Haizea-Llevant tool according to age (Drachler et al., 2007; Lopez-Boo et al., 2020). A quantitative coefficient for continuous variable analysis in the Haizea-Llevant tool was obtained by scoring the Delayed items as minus one point (−1) and Caution items as zero (0) and totaling the result. A Positive answer or performance in HLL is scored with one point if child’s performance is equal to or better than that of 50% or more of the standardization population for their age.

The CARE Booklet

Parents, mainly mothers to our case (55.8%), received a CARE booklet to be used as a screening report. The report consists of a mark over an icon (**Figure 3**), for which the parent or caregiver chooses *Sí* (“Yes”) if the skill or behavior was observed in interaction with the child, *No* if the skill or behavior was not observed in interaction with the child, or *No lo pude observar o creo que no lo puede hacer* (“I couldn’t observe it or I believe they can’t do it”) if the parent did not have an opportunity to observe if the skill or behavior were attainable by the child. The two options fall under the same question because the main intention with the booklet is the report of interactions, not recalls or beliefs about the children’s skills. The components of the CARE booklet keep the same dimensions but vary in the complexity of items between 24–35 months old and 36–47 months old. The content for 36–47-month-old children is the same as for 48–59-month-olds. The CARE instrument has 47 items in four domains comparable with the HLL observations: (a) personal-social (11 items), (b) language and logico-mathematical reasoning (20 items), (c) fine motor-adaptive (9 items), and (d) gross motor (7 items). It also includes an exploration of socio-cognitive development in context, in the use of Core Knowledge Systems (Kinzler and Spelke, 2007; Callaghan et al., 2011). The “Core Knowledge” components inquired with CARE are related to spontaneous and autonomous play, counting, geospatial orientation, age-pair interactions and outdoors activities. The Core Knowledge components used do not differ between each age-group booklet. The nominal classification and agreement analyses do not include the Core Knowledge components.

For nominal classification with the results in CARE, we followed the HLL scoring system, but included an arbitrary range for the not reported interactions when parents use the “I can’t observe it or I believe he/she can’t do it” option: if the child at least one Delay or at least two Cautions or at least four unanswered items (i.e., “I can’t observe it or I believe he/she can’t do it”) he/she was classified “At risk.” ‘No Delay’ answers or less



than two Cautions or ≤ 3 not answered items he/she would be classified 'Not at risk.' A quantitative coefficient for continuous variable analysis in CARE performance was obtained by scoring the Delayed items with -1 and Caution items with 0. A positive answer or performance in CARE was scored with 1 point.

Procedure

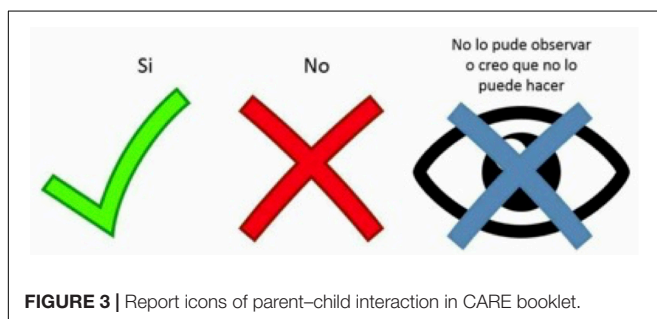
Children who screened positive for risk in a first screening, participated at a follow-up HLL screening at children's centers (CCs). The follow-up was performed by three trained assessors in an individual meeting with caregivers and children. During the second and final HLL screening, one of the assessors applied a survey to obtain sociodemographic information. Survey and screening application lasted less than 30 min. For children who screened positive in the initial session, a member of the research team contacted caregivers in the CC to administer the follow-up screen using HLL. A licensed psychologist then checked that assessors had completed all evaluations and proceeded to deliver a copy of the CARE booklet. Parents watched an instructional

2-min video on how to report children's activities using the CARE booklet. Families were instructed and directed explicitly to principal caregivers (**Table 1**) to carry out the activities and return the booklet as soon as possible but not less than 1 month after receiving it. After they had watched the video with the reporting instructions, the CARE booklet was delivered to the caregiver with the following items in a toy bag for each child: five wooden cubes, two hand puppets, a small plastic ball, one maraca, a preschooler's set of scissors, six crayons of different colors, and a pen with lid. Specific indications were given to parents to administer all items at home, and they were advised not to worry if their child did not complete them all. All children were screened in their primary language, Spanish.

The review board at the Faculty of Psychology (*Facultad de Psicología*) and the General Directorate of Research (*Dirección General de Investigaciones*) of the *Universidad de la Sabana* granted ethical approval for the study (Acta CAG #1517 of 19/11/2015). Permission for data collection was granted in agreement with the legal ruling of Resolution N° 008430 of 1993 of the *Ministerio de Salud de la República de Colombia* (Health Ministry of Colombia), which sets out ethical, scientific, technical and administrative norms for research activity with human participants. At the time of screening, parents were given an information sheet describing the larger original study. Consent for participation in the research project was indicated by completion of the sociodemographic survey, prior to inclusion in the current study.

Analysis

The analyses used average-based change statistics (ABCs), such as Cohen's d or Hays's ω^2 , to evaluate changes in distributions, and individual-based change statistics (IBCs), such as the



Standardized Individual Difference (SID) or the Reliable Change Index (RCI), to evaluate whether each case in the sample experienced a reliable change (Clifton and Clifton, 2019; Estrada et al., 2019). The standardization of measurement differences was used to calculate the net percentage change index [i.e., $100 \times (\text{CARE score} - \text{HLL score})/(\text{HLL score})$]. Primary analyses included mixed design analysis of variance (ANOVA), with data source (i.e., direct assessment using HLL, parental report using CARE) as a within-subjects factor and screening category group (i.e., “At risk” or “Not at risk”) as a between-subjects factor, to examine consistency between HLL and CARE in determining the developmental milestones reached. Separate mixed design ANOVAs were run for each developmental domain. The decision to use a mixed design ANOVA was based on the need to compare differences between groups split on two factors: a within-subjects factor in which all participants, serving as their own matched pair, were measured in two conditions (i.e., sources of information); and a between-subjects factor in which participants were classified separately based on DS. This analytic approach follows Miller et al.’s (2017) agreement study comparing direct testing and parent reports, while also allowing evaluation of the predictive quality of CARE booklet as a screening tool.

Secondary analyses included chi-square tests of agreement on individual matched pairs of items from both primary study measures, to determine agreement at the level of specific developmental milestones. In cases where assumptions of chi-square testing were violated due to small sample sizes (i.e., less than five cases in a contingency table cell), Fisher’s exact test was used.

Using the scoring procedures described above, interviewers’ direct observations with HLL and parental reports using CARE were scored by the author and checked independently by a licensed psychologist who was a research team member. Discrepancies in scoring were resolved in face-to-face meetings of the research team and compared against hard copies of the forms, and corrections were made on the forms. Demographic form data were entered into Microsoft Excel, uploaded to a drive-in cloud storage and checked using a double-data entry procedure.

Within our main results (i.e., participant recruitment and prevalence of developmental delay), the comparative analysis for CARE using parents’ report and direct observation included:

- (1) Effects of demographic variables (e.g., socioeconomic status) on overall agreement.
- (2) Effects of demographic variables on the various domain scores (personal-social, language and logico-mathematical reasoning, fine motor-adaptive, gross motor skills).
- (3) Overall agreement and congruence between the CARE report classification and interviewers’ direct screening classification (“At risk” or “Not at risk”), defined as the degree of correspondence between individuals’ judgments or ratings (Price et al., 2017). Inter-rater reliability (Cohen’s κ) was calculated and interpreted with the most accepted arbitrary ranges for Cohen’s κ (Landis and Koch, 1977): 0.00 – 0.20 indicates slight agreement, 0.21–0.40 fair agreements, 0.41–0.60 moderate agreement, 0.61–0.80

substantial agreement, and 0.81–1.00 indicates almost perfect agreement.

- (4) Screening classification (“At risk” or “Not at risk”) differences in development domain scores between HLL and parental CARE report. Differences in counting of total “No” answers in CARE reports and “Caution” items (i.e., an age-appropriate item is not passed) in HLL were analyzed. Also, differences were reported on domain scores (personal-social; language and logico-mathematical reasoning; fine motor-adaptive; gross motor skills) for both sources of data.
- (5) ROC curve area under the curve (AUC) analysis. The receiver operating characteristic (ROC) method is a commonly used paradigm in different medical and social areas to assess the performance of a diagnostic test (e.g., Schafer et al., 2014; Zanca et al., 2012). For the present study, our method requires values of two variables for each case: a truth variable (sometimes referred to as a ‘gold standard’) indicating the “At risk” status (HLL data) for each child and a decision variable indicating the CARE determination of “At risk” or “Not at risk.” The parent report in CARE is used to assign a single rating to each case (“At risk” or “Not at risk”). When the decision in CARE corresponds to the truth HLL direct observation status (“At risk”) it is called a true positive. When the decision in CARE does not correspond (i.e., “Not at risk”) to the truth HLL direct observation status (“At risk”) it is called a false negative. False positives correspond to a case when CARE reports an “At risk” condition but HLL indicates “Not at risk.” The ROC curve is a plot of true positive fraction in the sample (Sensitivity) and the complement of false positive fraction (Specificity) or $1 - \text{Specificity}$. When ROC uses non-parametric estimation for diagnostic test analyses (e.g., the Wilcoxon test), it is called an “empirical ROC” (Pepe, 2003). An empirical ROC has an empirical AUC. The area under the curve has a value between 0 and 1 showing the performance of the test (CARE), with higher values indicating better test performance and 0.5 indicating randomness. For small sample sizes, the empirical AUC may change dramatically due to small perturbations and differ significantly from the expected AUC (Ma et al., 2006). An alternative to the empirical AUC is the binormal AUC (Pepe, 2003). The binormal AUC is more stable than the empirical version for small sample sizes (Ma et al., 2006). In order to present comparable empirical AUC and binormal data, I report the nominal classification analysis using previous sensitivity and specificity calculation in a web page calculation tool (VassarStats: Website for Statistical Computation) and using quantitative indices for CARE and HLL classification to plot a binormal ROC curve (Eng, 2014).
- (6) Item-Level Comparison of Agreement for specific Domains. To determine agreement at the item level, a series of chi-square tests of agreement between parental reports and direct assessment was performed on individual matched item pairs. Inter-rater reliability (Cohen’s κ) and phi or Cramer’s V from the chi-square tests were reported

(Bakker and Wicherts, 2011). A Cramer's V parameter is used to compare the strength of association between any two cross-classification tables: a larger value for Cramer's V can be considered to indicate a strong relationship between variables, with a smaller value for V indicating a weaker relationship (Price et al., 2017).

- (7) Acceptability and feasibility analysis, which included six characteristics considered to influence implementation feasibility (Boggs et al., 2019): cultural adaptability, accessibility, training, administration time, geographical uptake, and clinical relevance and utility.

When necessary, in the following analyses, assumptions of normality, homogeneity of variances, and sphericity were met, and no significant outliers were identified in our sample. Otherwise, non-normal distribution of data was analyzed with non-parametric tools (i.e., the Kruskal-Wallis test or Mann-Whitney test). An alpha level of 0.05 was adopted for all statistical tests. All statistical analyses were conducted using IBM SPSS Statistics for Macintosh, Version 25.0 (IBM Corporation, 2017).

RESULTS

Prevalence of Developmental Delay

Using HLL, 75% of participants were classified "At risk" ($n = 39$). The CARE booklet reported that 71% ($n = 37$) of the sample qualified as "At risk" (Figure 4). Nominal classification analysis indicated that the sensitivity proportion was high (95%, corresponding to 37 out of 39 at-risk children), as was the specificity value (85%, corresponding to 11 out of 13 not-at-risk children). Also, the positive likelihood ratio ($LR+$) was 6.17 and the negative likelihood ratio ($LR-$) was 0.06.

Effect of Demographics on Overall Agreement

Analyzing the effect of demographic characteristics in overall agreement requires individual-based change statistics (IBCs) with the net percentage change index (NET). NET is calculated by $[100 \times (\text{CARE score} - \text{HLL score})/(\text{HLL score})]$. NET values indicate that the higher the difference score, the higher the probability of not agreement (Table 2). Also, negative values indicate lower score for the parental report in CARE compared to observation score using HLL (i.e., an underrated report by the parent). Differences between HLL and CARE report were higher in low SES (i.e., the second level) compared to very low SES homes. The medium-low SES was the only level at which the CARE score was lower than the HLL score.

One-way ANOVAs were then run to determine whether any sociodemographic variable had an effect on overall CARE and HLL score agreement. There was a main effect of SES on overall differences, $F(2,43) = 6.947$, $p = 0.002$, $\eta^2 = 0.12$. *Post hoc* analyses using the Bonferroni adjusted criterion for significance and t-test when significant differences were found, indicated that differences in scores were significantly higher in low SES compared with very low SES homes, $t(30) = -2.72$, $p = 0.011$,

$d = 0.72$, and with medium low SES, $t(31) = 2.98$, $p = 0.006$, $d = 0.81$.

No significant effect of other sociodemographic variables, including whether the child was a boy or a girl, was found on overall scoring differences between data sources (HLL vs. CARE) in the total sample.

Effect of Demographics on Domain Scores

Individual difference scores were calculated for analyzing the effects of demographic characteristics in every developmental domain assessed with HLL and CARE screening. The net percentage change index (NET) was calculated by subtracting each age-equivalent standardized individual CARE score from the age-equivalent standardized individual score in the corresponding developmental domain (Table 3).

Raw differences or standardized Individual Differences (SID) with negative values indicate lower score for the parental report in CARE compared to observation score using HLL (i.e., underrated report by parent). All medians with negative values indicate a central tendency with lower scoring in CARE report compared with HLL's scoring. Differences were higher in Personal-social and Gross motor domains for girls. Language and logico-mathematical reasoning and Fine motor-adaptive domains scorings has higher differences for boys. Working mothers had higher differences in Personal-social and Fine motor-adaptive for Employed status. Language and logico-mathematical reasoning and Gross motor domains scorings has higher differences for Unemployed status. Also, differences were higher in Personal-social domain for Medium low SES and in Language and logico-mathematical reasoning for Low SES (i.e., the second level). Fine motor-adaptive and Gross motor domains scorings have higher differences for Very low SES compared with other SES levels.

A Mann-Whitney test indicated a significant effect of working-mother status, with higher difference for employed (*Median* = -13.2) than unemployed mothers (*Median* = -11.7) on HLL and CARE scorings in the fine motor-adaptive domain, $U = 114.5$, $p = 0.02$, $r = 0.33$.

No significant effect of any other sociodemographic variables was found on developmental domains differences between data sources (CARE vs. HLL), suggesting that parents did not significantly differ in their ratings of child skills using CARE compared to direct testing with HLL in the total sample.

Overall Agreement Between Haizea-Llevant and CARE Screening Classification ("At Risk," "Not at Risk")

When comparing the classification outcomes of CARE booklet with the HLL, the overall agreement was 92% (by accuracy). Cohen's κ was calculated to determine if there was an agreement between the nominal screening classifications ("At risk" or "Not at risk") in HLL and CARE. There was almost perfect agreement between the two classifications data, $\kappa = 0.810$ (95% CI -0.973, -0.988), $p < 0.0001$.

CARE parent report vs. Haizea-Llevant screening

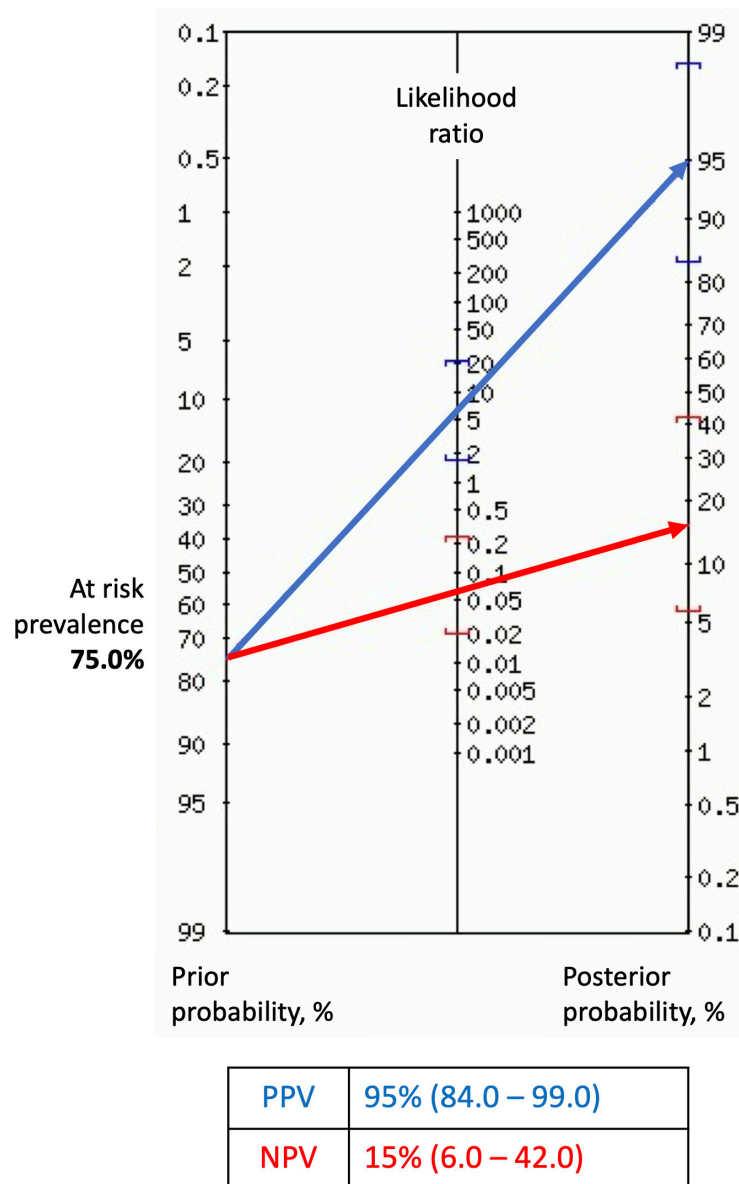


FIGURE 4 | Fagan's nomogram showing probability of children At risk after parents report using CARE booklet. Probabilities were calculated based on the screening with Haizea-Llevant table (HLL). Positive At risk diagnosis (blue arrow) refers to typical or non-specific appearance, and Not at risk diagnosis (red arrow) to atypical or negative appearance in CARE. Precision is given as 95% confidence interval. Risk prevalence is derived from the number of At risk positive and Not at risk participants after screening with HLL. LR+, positive likelihood ratio; LR-, negative likelihood ratio; NPV, negative predictive value; PPV, positive predictive value. Diagnostic test calculator (version 2010042101). Copyright (c) 2002-2006 by Alan Schwartz < alansz@uic.edu >.

Screening Classification (“At Risk,” “Not at Risk”) Differences in Delay and Caution Items Between Haizea-Llevant and CARE

Table 4 presents descriptive statistics of overall performance on items (i.e., Delays and Cautions) and nominal classification (i.e., “At risk” or “Not at risk”) using HLL and parents’ reports using CARE. In the HLL reports, more items were reported

as Cautions than Delays. The same was true for CARE reports in “Not at risk” participants. Contrary, Delays were four times more likely to be reported in “At risk” children when using the CARE report.

A Mann-Whitney tests indicated a significant difference in HLL observations, such that the “At risk” group presented a greater number of Caution items (*Median* = 3) than the “Not at risk” group (*Median* = 1), $U = 66.0$, $p < 0.001$, $r = 0.56$. Similarly, “At risk” children presented a greater number of Delay items

TABLE 2 | Raw and net percentage change index (NET) for overall scoring differences between Haizea-Llevant (HLL) and CARE.

| SES | n(%) | Haizea-Llevant overall (raw) scoring | | CARE overall (raw) scoring | | HLL minus CARE overall NET ⁺ difference | |
|----------------------|-----------|--------------------------------------|------|----------------------------|------|--|-------|
| | | M | SD | M | SD | M | SD |
| Level 1 – Very low | 13 (25) | 0.67 | 0.11 | 0.68 | 0.08 | 3.41 | 16.21 |
| Level 2 – Low | 19 (36.5) | 0.67 | 0.19 | 0.78 | 0.11 | 25.94 | 40.41 |
| Level 3 – Medium low | 14 (26.9) | 0.72 | 0.11 | 0.70 | 0.13 | –0.57 | 23.31 |
| No data | 6 (11.5) | | | | | | |

⁺ $100 \times (\text{CARE score} - \text{HLL score})/(\text{HLL score})$.

TABLE 3 | Median and data spread (Interquartile range-IQR) for the Net percentage change index (NET) between scores for Haizea-Llevant (HLL) and CARE report by developmental dimensions.

| | Personal-social domain | | Language and logico-mathematical reasoning | | Fine motor-adaptive domain | | Gross motor domain | |
|------------------------------|------------------------|------|--|------|----------------------------|------|--------------------|------|
| | Median | IQR | Median | IQR | Median | IQR | Median | IQR |
| Sex | | | | | | | | |
| Male | –12.7 | 18.9 | –11.2 | 16.9 | –15.2 | 21.4 | –10.0 | 27.6 |
| Female | –12.8 | 12.5 | –8.5 | 10.8 | –11.4 | 13.4 | –13.3 | 11.3 |
| Working mother status | | | | | | | | |
| Employed | –13.5 | 19.6 | –8.6 | 16.6 | –13.2 | 19.4 | –14.6 | 23.2 |
| Unemployed | –9.4 | 13.1 | –16.6 | 18.4 | –11.7 | 28.0 | –19.9 | 38.9 |
| SES | | | | | | | | |
| Level 1 – Very low | –15.4 | 27.0 | –7.7 | 15.4 | –15.2 | 9.7 | –16.1 | 9.4 |
| Level 2 – Low | –8.2 | 13.1 | –16.8 | 15.6 | –8.1 | 32.6 | –9.3 | 40.7 |
| Level 3 – Medium low | –16.5 | 28.8 | –8.6 | 9.0 | –8.3 | 21.6 | –6.9 | 35.5 |

TABLE 4 | Delays and Cautions for nominal classification groups using Haizea-Llevant (HLL) and CARE.

| | n (%) | Items in Delay | | Items in Caution | |
|--------------------------|-----------|----------------|-----|------------------|-----|
| | | Median | IQR | Median | IQR |
| HLL-Observation | | | | | |
| At risk | 39 (0.75) | 1.0 | 2.0 | 3.0 | 3.0 |
| Not at risk | 13 (0.25) | 0.0 | 0.0 | 1.0 | 0.0 |
| Using CARE report | | | | | |
| At risk | 39 (0.75) | 4.0 | 3.5 | 1.0 | 4.5 |
| Not at risk | 13 (0.25) | 0.0 | 1.0 | 1.0 | 1.0 |

(Median = 4) than the “Not at risk” group (Median = 0), $U = 85.5$, $p < 0.001$, $r = 0.50$.

Screening Classification (“At Risk,” “Not at Risk”) in Development Domain Scores for Haizea-Llevant and CARE

Standardized individual scores were calculated for analyzing developmental dimensions (i.e., Personal-social domain) and nominal classification (i.e., “At risk” or “Not at risk”) using both HLL and CARE (Table 5). Differences were greater in HLL classification in the personal-social and language and logico-mathematical reasoning domains for “Not at risk” children. Also, same children (HLL classification: “Not at risk” children) had a higher CARE report scoring than their HLL score in the gross motor domain. Fine motor-adaptive scorings had higher differences for “At risk” children classified using HLL observation. Greater differences with higher CARE report scoring than HLL score were seen for “Not at risk” children in all domains.

A Mann-Whitney test indicated that scores on the CARE report in the personal-social domain were lower for the “At risk” group (Median = 0.7) than for the “Not at risk” group (Median = 1.0), $U = 82.5$, $p = 0.001$, $r = 0.52$. No significant difference was found between “At risk” or “Not at risk” groups on personal-social domain scores for direct testing with HLL. Comparing scores in language and logico-mathematical reasoning, using a Mann-Whitney test, indicated that on CARE report scores were lower for the “At risk” group (Median = 0.7) than for the “Not at risk” group (Median = 1.0), $U = 74.0$, $p = 0.001$, $r = 0.53$. No significant difference was found between “At risk” or “Not at risk” groups on language and logico-mathematical domain scores for direct testing with HLL. Also, a Mann-Whitney test indicated that score in fine motor-adaptive domain on CARE report was lower for the “At risk” group (Median = 0.8) than for the “Not at risk” group (Median = 1.0), $U = 118.5$, $p = 0.01$, $r = 0.42$. No significant difference was found between “At risk” or “Not at risk” groups on fine motor-adaptive domain scores for direct testing with HLL in the total

TABLE 5 | Median and data spread (Interquartile range-IQR) for the Net percentage change index (NET) between scores for Haizea-Llevant (HLL) and CARE report by developmental dimensions.

| | <i>Personal-social domain</i> | | <i>Language and logico-mathematical reasoning</i> | | <i>Fine motor-adaptive domain</i> | | <i>Gross motor domain</i> | |
|--------------------------|-------------------------------|------|---|------|-----------------------------------|------|---------------------------|------|
| | Median | IQR | Median | IQR | Median | IQR | Median | IQR |
| HLL-Observation | | | | | | | | |
| At risk | 0.20 | 0.98 | 0.00 | 1.74 | -0.15 | 1.51 | 0.00 | 1.46 |
| Not at risk | -0.29 | 2.83 | -0.59 | 1.25 | 0.13 | 1.85 | 0.73 | 0.00 |
| Using CARE report | | | | | | | | |
| At risk | -0.30 | 1.12 | -0.22 | 1.06 | -0.02 | 0.00 | 0.05 | 0.00 |
| Not at risk | 0.89 | 0.00 | 1.03 | 0.43 | 0.81 | 0.00 | 0.73 | 0.00 |

sample (data not shown). Score in gross motor domain on CARE report, a Mann-Whitney test, indicated that was lower for the “At risk” group (*Median* = 0.80) than for the “Not at risk” group (*Median* = 1.0), $U = 110.5$, $p = 0.01$, $r = 0.45$. Likewise, scores in gross motor domain on direct testing with HLL was lower for the “At risk” group (*Median* = 0.75) than for the “Not at risk” group (*Median* = 1.0), $U = 72.5$, $p = 0.05$, $r = 0.30$.

Receiver Operating Characteristic Curve: Area Under the Curve

When performing an empirical ROC-curve analyses in the total sample ($n = 52$), the area under the curve (AUC) is 0.894 (Trapezoidal Wilcoxon area) with a higher Youden index of 0.860 (Supplementary Table 1). Otherwise, a binormal ROC curve (Figure 5) uses quantitative index for CARE and HLL classification as a truth variable indicating the “At risk” status for each child. The Area under the fitted curve (A_z) in the binormal curve is 0.899.

Youden J indexes (Supplementary Table 1) are reported because they indicate the maximum potential effectiveness of CARE scoring, and act as a common summary measure of the ROC curve (Ruopp et al., 2008).

Item-Level Comparison of Agreement for Specific Domains

Given the small group sizes when the sample was split by demographic variables, item level analyses were conducted on the full sample instead of separately for each screening group. Table 6 shows the mean proportions of correct items in the HLL and CARE reports. An important aspect to note is the asymmetry in the number of participants due to the application of HLL to specific ages and the delivery of CARE to the general sample. After descriptive data, the agreement at the item level was determined with a series of chi-square tests, performed on individual matched item pairs across HLL and CARE scores and developmental dimensions.

Several chi-square tests indicated, overall, somewhat mixed item-level agreement findings for every domain. The proportion of items with significant agreements was higher in personal-social (7 out of 11: 63%) and language and logico-mathematical reasoning (14 out of 20: 70%) than the proportions in fine motor-adaptive (5 out of 9: 55.5%) and gross motor skills (3 out of 7: 42.8%). However, nearly all scores for items accrued

in one quadrant of the chi-square contingency table. Under that condition there are key limitations to adequate interpretation for Kappa values for agreement between data sources. That is a reason to report Cramer's V (Gingrich, 2004), which is used to compare the strength of association between any two cross-classification tables. Tables which have a larger value for Cramer's V can be considered to have a strong relationship between the variables, with a smaller value for V indicating a weaker relationship (Gingrich, 2004).

Personal-Social Domain

For items assessing personal-social domain (e.g., “Help in house”), there was more significant agreement than

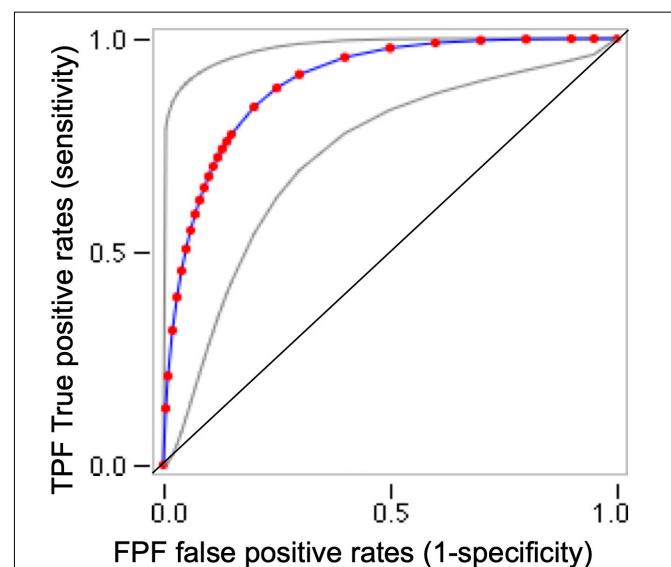


FIGURE 5 | Receiver operating characteristic (ROC) binormal curve for CARE and Haizea-Llevant classification for the total sample ($n = 52$). This ROC curves plot use web-based calculator for ROC curves (<http://www.jrocf.it.org>). Gray lines indicate 95% confidence interval of the fitted ROC curve. ROC analysis plot for each possible cut-off points of the relevant CARE scale, the true-positive proportion (sensitivity = 95%) against the false-positive proportion (1– specificity). A perfect test would have an area under the curve (AUC) of 1 and the curve would pass through the upper left corner of the plot (100% sensitivity, 100% specificity). In this study, Trapezoidal (Wilcoxon) area/AUC = 0.89 ($SE = 0.04$) and the Area under the fitted curve (A_z) = 0.90 ($SE = 0.052$).

TABLE 6 | Media and standard deviation (SD) for assertive observation or reports in Haizea-Llevant (HLL) and CARE by items in developmental dimensions.

| | HLLlevant | | | CARE | | |
|---|-----------|----------|-----------|----------|----------|-----------|
| | <i>n</i> | <i>M</i> | <i>SD</i> | <i>n</i> | <i>M</i> | <i>SD</i> |
| Personal-social domain | | | | | | |
| Help in House | 4 | 1.00 | 0.00 | 9 | 0.67 | 0.73 |
| Feed doll | 7 | 0.86 | 0.76 | 9 | 0.89 | 0.67 |
| Remove Garment | 12 | 1.00 | 0.00 | 9 | 0.89 | 0.67 |
| When he or she play with dolls, he/she performed a play like a script or short tale with their dolls or toys? | 17 | 0.94 | 0.49 | 52 | 0.92 | 0.36 |
| Put on clothing | 30 | 0.56 | 1.00 | 52 | 0.77 | 0.74 |
| Did he/she suggest or show when need to go to the toilet? | 17 | 1.00 | 0.00 | 50 | 0.88 | 0.27 |
| Did he/she answer if he or she is a boy or a girl? | 30 | 0.78 | 0.86 | 52 | 0.90 | 0.50 |
| Dress, no help | 26 | 0.41 | 1.02 | 52 | 0.71 | 0.85 |
| Did he/she play with an adult using hand puppets? | 31 | 1.00 | 0.68 | 52 | 0.87 | 0.41 |
| Prepare cereal (In Spanish this item is open to more food than cereals) | 24 | 0.64 | 0.95 | 43 | 0.84 | 0.67 |
| Draw a person | 16 | 0.44 | 0.91 | 43 | 0.53 | 0.95 |
| Language and logico-mathematical reasoning | | | | | | |
| Name __ Pictures (6 pictures) | 5 | 0.87 | 1.10 | 9 | 0.67 | 0.88 |
| Know 2 actions | 5 | 0.63 | 1.10 | 9 | 0.78 | 0.71 |
| Combine words | 5 | 0.40 | 1.10 | 9 | 0.56 | 0.87 |
| Name __ Pictures (5 pictures) | 9 | 0.56 | 1.05 | 9 | 0.89 | 0.33 |
| Use of 3 Objects | 10 | 0.40 | 0.97 | 9 | 0.89 | 0.33 |
| Speech half understandable | 12 | 0.40 | 0.90 | 26 | 0.89 | 0.33 |
| Did he/she point the dog correctly? (memorize an image) | 19 | 0.70 | 0.96 | 35 | 0.85 | 0.59 |
| When he or she speaks use pronouns? | 29 | 0.28 | 0.94 | 9 | 0.97 | 0.17 |
| Did he/she count aloud two consecutive numbers? | 27 | 0.43 | 1.02 | 52 | 0.79 | 0.71 |
| Name __ Pictures (10 pictures) | 33 | 0.68 | 1.01 | 52 | 0.96 | 0.19 |
| Did he/she use "to be" in a phrase? | 33 | 0.30 | 1.00 | 52 | 0.90 | 0.50 |
| Pick longer line | 38 | 0.42 | 1.01 | 52 | 0.90 | 0.55 |
| Speech all understandable | 37 | 0.51 | 0.99 | 51 | 0.85 | 0.62 |
| Identify colors | 36 | 0.50 | 0.96 | 43 | 0.79 | 0.82 |
| Did he/she realize no-connected actions? | 39 | 0.63 | 0.97 | 43 | 0.88 | 0.55 |
| Name colors | 27 | 0.54 | 0.90 | 43 | 0.79 | 0.68 |
| Opposites – morning/afternoon | 23 | 0.36 | 0.93 | 43 | 0.79 | 0.68 |
| Did he/she tell stories? | 16 | 0.62 | 0.25 | 43 | 0.63 | 0.92 |
| Did he/she repeat a complete phrase? | 12 | 0.41 | 0.51 | 43 | 0.67 | 0.83 |
| Did he/she recognize numbers (Arabic writing numerals)? | 12 | 0.42 | 0.52 | 43 | 0.56 | 0.92 |
| Fine motor-adaptive domain | | | | | | |
| Put Block in Cup | 6 | 0.94 | 0.00 | 9 | 1.00 | 0.00 |
| Tower of 4 cubes | 9 | 0.62 | 0.00 | 9 | 1.00 | 0.00 |
| Thumb-finger grasp (grab a pencil) | 16 | 0.54 | 1.03 | 52 | 0.88 | 0.46 |
| Copy a circle | 30 | 0.00 | 1.02 | 52 | 0.87 | 0.61 |
| Did he/she imitate a bridge with 3 cubes? | 37 | 0.00 | 1.01 | 52 | 0.87 | 0.57 |
| Did he/she fold a paper sheet? | 30 | 0.74 | 0.82 | 44 | 0.73 | 0.69 |
| Did he/she use scissors to cut a paper sheet? | 26 | 0.59 | 0.98 | 44 | 0.77 | 0.64 |
| Copy a square | 19 | 0.53 | 1.01 | 44 | 0.64 | 0.82 |
| Did he/she imitate a door with 5 cubes? | 19 | 0.79 | 0.84 | 44 | 0.73 | 0.69 |
| Gross motor domain | | | | | | |
| Walk down steps | 4 | 0.58 | 0.00 | 9 | 1.00 | 0.00 |
| Kick ball forward | 4 | 0.58 | 1.00 | 9 | 1.00 | 0.00 |
| Broad jump | 17 | 0.79 | 0.87 | 52 | 0.85 | 0.58 |
| Balance Each Foot 5 s | 28 | 0.00 | 0.92 | 52 | 0.75 | 0.75 |
| Jump up | 29 | 0.25 | 0.82 | 52 | 0.79 | 0.64 |
| Did he/she jump backwards? | 22 | 0.76 | 0.46 | 52 | 0.69 | 0.66 |
| Balance each foot 1 s | 18 | 0.79 | 0.57 | 44 | 0.75 | 0.69 |

non-agreement between parental report and direct testing (Supplementary Table 2). However, on some items measuring-agreement continuity is expected, because some activities will use the same objects in a trajectory of increasing complexity in interactions with adults or peers. Items like “Feed doll” and

“When he or she plays with dolls, he/she performed a play like a script or short tale with their dolls or toys?” or “Did he/she play with an adult using hand puppets?” are examples of the expected trajectory. The expected trajectory apparently requires more complex developmental skills that affect the agreement

level. Another example is “Remove garment” and “Put on clothing” or “Dresses, without help.” For those items, parents mostly reported that the child had the skill, but it was not seen on direct testing. Finally, a significant disagreement ($\kappa \leq 0$) between CARE and HLL direct testing was found in “Did he/she suggest or indicate needing to go to the toilet?”, showing that this particular behavior was more often seen in direct assessment than reported by parents.

Language and Logico-Mathematical Reasoning

For items assessing language and logico-mathematical reasoning skills (e.g., “Combine words”), there were more items in significant agreement than items with non-agreement between parent report and direct testing (**Supplementary Table 3**). However, as in the personal-social domain, there were items where measuring-agreement continuity was not obtained, e.g., “Did he/she count aloud two consecutive numbers?” and “Did he/she recognize numbers (Arabic numerals)?”. Also, perceptual and contextual discrimination skills were not in agreement (i.e., parents reported that the child could “Pick longer line” and recognize “Opposites - morning/afternoon” more often than seen on direct assessment). Likewise, some expressive language items had no significant agreement (i.e., “Did he/she use ‘to be’ in a phrase?”; “Did he/she repeat a complete phrase?”).

Fine Motor-Adaptive Domain

For items assessing fine motor-adaptive skills (e.g., make a “Tower of four cubes”), there was almost the same number of items in significant agreement than those without significant agreement between parent report and direct testing (**Supplementary Table 4**). However, as with previous domains, there were items where measuring-agreement continuity was not obtained (i.e., “Tower of four cubes” vs. “Did he/she imitate a bridge with three cubes?”, and “Copy a circle” vs. “Copy a square”).

Gross Motor Domain

For items assessing gross motor domain (e.g., making a “Wide jump”), there were more items with no significant agreement than items with significant agreement between parent report and direct testing (**Supplementary Table 5**). As in previous domains, there were items where measuring-agreement continuity was not obtained (i.e., “Wide jump” and “Jump up”).

Acceptability and Feasibility

The rating criteria in Boggs et al. (2019) for mentioned characteristics in screening tools were applied to the CARE reports. Validity and reliability analysis was presented in previous sections. According to Boggs et al. (2019), CARE presented several characteristics in rating levels between 0 and 3, indicating a good consideration for scalable studies (**Table 7**).

DISCUSSION

The CARE booklet featured in this study aims to monitor and support parents’ interactions for enhancing children’s

development and identify developmental difficulties. The previous phases of this study include the conceptualization and consolidation of CARE components related to the Haizea-Llevant DS table (HLL). The monitoring component of CARE is central to the current study reported here, in particular an examination of its sensitivity and specificity in a small sample of vulnerable families in Colombia. The sample of families and children recruited from a community children’s center in Colombia’s capital, Bogotá, was similar to those for which similar screening tools are designed and standardized in LMIC populations (Faruk et al., 2020).

Firstly, a positive characteristic of CARE is in the level of engagement shown for a measurement tool relating to a cognitive intervention. Following a meta-analysis for commitment of parental involvement (Haine-Schlagel and Escobar-Walsh, 2015), completion of tasks in cognitive interventions had a range of 19–89% in participants. The effective users of the CARE booklet in this study were the 85.2% of receivers who used it for 1 month at home. The high level of CARE report use has considerable positive implications for the whole monitoring, screening and surveillance cycle to track a child’s developmental progress (Faruk et al., 2020), known as the detection-intervention-prevention continuum.

Second, concerning the prevalence of developmental delay, our procedure to recruit participants after a first screening may have affected the high level of delay found (75%), raising concerns for more wide-ranging recruitment in an experimental field procedure using CARE as a screening tool. However, recent studies reported low delay prevalence in DS (Ozturk-Ertem et al., 2019) and the higher prevalence in our study must be interpreted with caution. If excluding participants to receive the CARE booklet after first screening is a recruitment bias, it is an opportunity for methodological improvement since several barriers to the identification of developmental delay using tools adapted for LMIC have recently been reported (Faruk et al., 2020). Indeed, other screening studies include samples that did not share comparable sociodemographic characteristics to our participants, such as lower socioeconomic status (Murphy et al., 2020). According to the expressed aims of the current study, next discussions comprehend the specific results.

Consistency Between CARE and Haizea-Llevant Classification and Scores in Developmental Domains

Overall, the results suggest that parental observation of different child abilities reported in the CARE booklet did not differ significantly from direct assessment using HLL, and results were generally stable across screening classification groups (i.e., overall agreement by accuracy: 92%). Also, the effects of demographic variables on agreement between parent report and direct assessment of child are fundamental for decisions on future research and interventions after the COVID-19 pandemic. Differences for lower socioeconomic status and working-mother status indicated a need for better tracking of interactions related to parenting employment and individual developmental trajectories when those demographic conditions

TABLE 7 | CARE characteristics according to early child development measurement tool accuracy and feasibility for use in routine programs criteria by Boggs et al. (2019).

| | Boggs level description | Observation about CARE |
|---|--|---|
| Cultural adaptability, Rating: 3 | Easy modification of items, materials and procedures. | All items have a particular space for annotations a personalize described instructions or activities. The modification of items, materials and procedures will be fitted according inhouse context. Pictures and words are widely understood for specific participants with low academic level. |
| Accessibility, Rating: 2 | Tool, administration, scoring and interpretation, adaptation and training resources all available open access online with no intellectual property restrictions, minimal cost to tool and/or equipment (\leq US\$10 per child), no app available. | CARE is online available at https://monitoreoencasa.weebly.com/The toys and materials delivered with the printed booklet cost less than 7 GBP per child. |
| Training, Rating: 3 | Brief (\leq 1 h), minimal (i.e., non-specialist worker can train non-specialist worker), no certification requirement. | Parents only received a less than 3 min video instruction (https://youtu.be/Y5864iGCvG8); research team are undergraduate students and do not receive specialized instruction for cooperation or answer questions coming from parents. |
| Administration time, Rating: 2 | > 15 to \leq 30 min, minimum to moderate scoring. | CARE is planned to apply at home. A direct question about accumulated time when the booklet is returned to research team indicates less than an hour throughout a 1 month. |
| Geographical uptake, Rating: 0 | Used in one country only. | Only used in Colombia. |
| Clinical relevance and utility, Rating: 3 | Easy interpretation, clear threshold for action and structure for counseling response and contextually appropriate referral. | CARE is intended to use it as referral for clinical surveillance and motive observations an interaction between caregivers and children at home. All individuals had a one-page results, as a guide for educative action and understandable by caregivers and CC workers in the individual report returned as feedback to participants. |

are present in LMIC populations (Campaña et al., 2020). Language and mathematical reasoning and fine motor skills were the two skill areas most affected by SES conditions in our data, in common with previous studies of early childhood (Justice et al., 2019). Some barriers connected with caregivers serving as informants of their own interactions' quality relate to parental distress around parent-child interactions. CARE DS might diminish parental stress or other contingent conditions associated with dysregulated parent-child interactions and reported in vulnerable or impoverished conditions (Justice et al., 2019). However, SES is not defined solely by economic poverty, and more research is need in order to clarify the issue of scarcity in child-parent interactions (Guan et al., 2020).

Altogether, these findings suggest that both CARE reports and direct testing are appropriate forms of child DS. However, this study has an advantage over other comparisons with agreement analyses, including Miller et al. (2017): 100% of items in the parental reports (the CARE booklet) were comparable with the items included in the direct screening measurement. Indeed, Miller et al. (2017) only compared 12 out of 381 items (3.15%) for the Vineland Adaptive Behavior Scales (Survey Interview Form; Sparrow et al., 2005) and 12 out of 91 items (13.2%) for the Mullen Scales of Early Learning (Mullen, 1995). The good agreement shown in our results suggests that parents are generally reliable reporters of child abilities. When comparing agreement between "At risk" classification and scores on CARE and HLL (see Tables 4, 5), across the domains of personal-social skills, language and logico-mathematical reasoning, fine motor-adaptive and gross motor skills, CARE demonstrated

discriminatory potential that was as good as that provided by the HLL direct observations.

In particular, while HLL is a better detector for Cautions, CARE demonstrated better discrimination for Delays. Furthermore, all developmental domains had differences in nominal classifications in the "At risk" and "Not at risk" groups using CARE, but only in the gross motor skills dimension using HLL. A next step in the optimal design process for CARE should be a comparison with other tools in order to establish wide discriminatory characteristics in a Field Testing-Analysis-Revision framework (Nadeem et al., 2016).

Item Level Consistency Between CARE and Haizea-Llevant

Overall, the proportion of items in agreement were higher for personal-social and for language and logico-mathematical reasoning compared to the proportions for fine motor-adaptive and gross motor skills. The obvious answer to explain this discrepancy would be the time dedicated to observation of interactions. CARE gives parents 1 month to screen their children constantly on four developmental dimensions. Unfortunately, an explicit limitation is in the lack of analysis for any difference regarding the time it takes for parents to complete the CARE booklet. That means a limitation in determining the effect of the whole time dedicated to use and return CARE, as it could be done in a day, during a week or over the whole month. However, these long-lasting observations with the screening activities in CARE relating to fine motor-adaptive and gross motor skills might increase the disagreement with the short-term observations using

HLL, given the accumulation of time and opportunities for reporting motor interactions at home. Otherwise, a significant disagreement ($\kappa \leq 0$) between CARE and HLL direct testing was found in “Did he/she suggest or indicate needing to go to the toilet?”, with this particular behavior more often seen in direct assessment than reported by parents. The autonomy levels expected in the test environment are different in the Children’s Center compared to the child’s home. Also, such items will be subject to parents’ interpretation according to the cultural context (Schiari et al., 2021). In this specific case, the lack of autonomy assigned to going to the toilet, and other social items, could result from parents assuming that a child cannot perform age-appropriate tasks without having actually observed these in detail at home (Miller et al., 2017). CARE screening might demand attention to behaviors, skills and performances that routinely are included in at-home interactions and excluded in the report. The attentional demands of routine interactions between parents and children were recently included in an analysis of associations between high levels of cognitive stimulation in the home and increased screening scores for children in low-SES conditions (Slemming et al., 2021). Specifically, they analyzed this under the so-called “standard model” of consecutive *knowledge* → *stimulation* → *development* (Bornstein, 2015; Britto et al., 2017; Cuartas et al., 2020).

The *knowledge* → *stimulation* → *development* ($K \rightarrow S \rightarrow D$) model acts like a “cascade” of processes and outcomes, involving parenting attributions and supportive parenting, and concluding in the child’s externalizing behavior. In the $K \rightarrow S \rightarrow D$ model, the testing of any particular child’s skills by observation has specific challenges for parents and even for professional experts in child development, despite their favorable knowledge and attitudes (Jain et al., 2021) and appropriate healthcare organizational setup (Sheeran et al., 2020). Child non-compliance reduced attention and interest in calls for interaction, and the unfamiliar framework for direct reports at home might affect the success of testing. Recent research confirms the relevance of responsive parental behavior and child’s interactive engagement for positive developmental trajectories in children with significant cognitive and motor developmental delay (Van Keer et al., 2020). The level of attention from parents, and the initiation of interactions by children, might explain why the frequency, continuity and quality of interactions at home affect positive parental reports when interaction is not complex, but disagrees with external observation when complexity in interactions is higher and is not capable of full reporting through the screening measurements. In our data, the disagreement levels were specifically noted in fine and gross motor skills (i.e., proportion of items without significant agreements: 57.2%), as we expected and was suggested before by Miller et al. (2017).

Moreover, the $K \rightarrow S \rightarrow D$ model implies that parents might recall whether a skill milestone had effectively been reached, before confirming this through observation. If the CARE delivery is not enough for changing parental knowledge of stimulating interactions and consequently affecting children’s outcomes, a pre-post study might indicate the need for a new design, beyond CARE delivery as an intervention with screening tools.

Diagnostic Characteristics and Performance of CARE as a Tool for Developmental Screening

Receiver operating characteristic analysis results indicated that CARE is a satisfactory tool for screening diagnostics and might help to build a quantitative index for better and faster classification of an “At risk” status in children aged 24–59 months. Our data offers complete diagnostic performance for a screening tool, surpassing the limitations of other tools designed and developed in LMIC (Faruk et al., 2020), such as the Child Language Test in Phonology, Vocabulary, Fluency and Pragmatics (ABFW), the Developmental Assessment Scales for Indian Infants (DASII), and the Rapid Neurodevelopmental Assessment (RNDA; Juneja et al., 2012; Khan et al., 2013; Dias et al., 2020). There is no ROC analysis of ABFW, DASII or RNDA to compare with our data. However, the sensitivity and specificity (95 and 85% respectively) of CARE were higher than for another tool validated against the Denver Developmental Screening Test, namely, the Trivandrum Developmental Screening Chart (TDSC). The TDSC had an overall sensitivity and specificity of 66.7 and 78.8%, respectively. The diagnostic characteristics of CARE are highly trustworthy compared to other screening tools designed for long observation periods by parents. However, due to the limitations set out in the next section, we cannot say that CARE might be better than the Guide for Monitoring Child Development (GMCD) or other tools targeted at early ages or specific developmental domains, such as social-emotional or self-help subscales (Faruk et al., 2020).

Pilot Validity of CARE for Research and Intervention With Institutional Community Participants

The CARE booklet, and other screening tools administered by parents, might act like home-based records (HBRs). Such records do not replace clinical or scientific intervention, but can run in parallel with other existing or subsequent screening tools for optimal health and educational system interventions (Mahadevan and Broadus-Shea, 2020). The CARE booklet shows similar conditions for delivery as HBRs, with rigorous reliability and agreement results. Also, CARE content and design had enough cultural adaptability to follow the Nurturing Care Framework and could be administrated in programs like FAMI for rural families in Colombia (Milner et al., 2019). Following the standards of Boggs et al. (2019) for screening tools, the accessibility of CARE might be diminished by the fact that there is no digital app for it available. However, this might not be true for families with lower resources or in some geographical regions, who may not access the internet. A first step considering the relevance of Boggs et al. (2019) but forgetting the focus on vulnerable and limited resources for families in poverty is in an online information-delivery through a beta webpage with a digital version of CARE¹. The availability of CARE in electronic format limits the delivery for the focused families in the present study. However, it will contribute to even easier access

¹<https://monitoreoencasa.weebly.com/>

and optimal conditions for training and administration time in families and health systems having non-limited connection or access to the internet.

Finally, as a preliminary conclusion, CARE may be an efficient, cost-effective screening instrument for children between aged 24–59 months who are at risk of not reaching all their cognitive potential because of social and economic limitations. The clinical relevance and utility of the accurate and efficient classification obtained with tools like CARE might be successfully included in health systems and surveillance routines for DS in the detection of delay, and can be useful for identification and electronic records as well (Vitrikas et al., 2017; Gellasch, 2019). Developmental monitoring and screening processes in LMIC should use tools like CARE for detecting and increasing early intervention referrals, assessments and eligibility for the children who need it most (Barger et al., 2018; Goldfeld and Yousafzai, 2018). CARE not only shows the desired sensitivity-specificity values, but also provides information on cultural adaptation with respect to the communities that use Children's Centers for vulnerable families in Colombia. The reported diagnostic and screening characteristics also most likely resulted in the high level of acceptance of the screening process (75.1%), which is crucial for the success of a large-scale surveillance program. However, attention to the limitations of this study and the possibility for further research is needed to evaluate its potential for population screening and monitoring, and its cost-effectiveness as a public health measure.

Limitations in CARE Screening and Diagnostics Characteristics

The lack of data about the clinical status of parents using CARE helps to maintain the consideration of parental discrepancy in reports as an essential source of information, given the assuming norm that parents are uniquely positioned to observe and interact with children in various situations at home (Bennetts et al., 2016; Miller et al., 2017; Jeong et al., 2019). However, the results of the item analysis require an explanation of certain disagreements and inconsistencies. The data appear overall to have no systematic pattern of disagreement in the consideration of items by domains (i.e., proportion of items with significant agreements, personal-social: 63%, language and logico-mathematical reasoning: 70%, fine motor-adaptive: 55.5%, gross motor skills: 42.8%), but some disagreements (e.g., “Copy a circle”: $\kappa = 0.015$, $p = 0.72$; “Copy a square” $\kappa = 0.125$, $p < 0.01$) show a truncated continuity in the screening process by parents when the nature of the activities increases the complexity in some domains. The $K \rightarrow S \rightarrow D$ model explain the probability of memory and recall use for parent's report, but do not resolve this issue in future and scalable applications of CARE. As indicated before, this a pilot phase of CARE for optimizing the design following the components of Nadeem et al. (2016) and several other limitations in the present study might be addressed before subsequent field testing.

Also, our standardized DS tool, the HLL has its own limitations. First, the last reported use and correction was normed a decade ago (Rivas et al., 2010) and it is thus less up-to-date than other early DS tools (Boggs et al., 2019). Second,

like any other screening test, CARE only allows for a ‘snapshot’ of a child at one time point, limiting the ability to capture the full range of a child's functioning. The CARE snapshot might lead to interpreting a false classification or disagreement at item level (compared to the HLL observation) as “parental error” (Miller et al., 2017, p. 12). Miller et al. (2017) argued that it cannot be systematically ascertained whether a child's behavior during the evaluation was typical of his or her home behavior. An alternative to the “error” explanation is a hypothesis related to the effects of the psychology of scarcity (Shah et al., 2012, 2015, 2018; Camerer et al., 2018). This argument might be called the “scarcity of parental interactions” argument as opposed to the error argument (Miller et al., 2017). For the other kind of disagreements, “when a parent reports that a child has a skill, yet the skill is not seen on direct assessment” (Miller et al., 2017, p. 12), parents might use two strategies to report using CARE: (a) recall or memory of interaction events, and (b) direct subsequent observations of their interactions with children. A limitation on analyzing these disagreements is in the lack of more invasive research and evaluation techniques in this study, with a clear suggestion of including home-visit observations or home-recorded videos.

Limitations in the Study Design and Further Studies

Using CARE as a screening tool have the potential to activate alerts for early cognitive delay that reassure clinicians and families of further specialized and controlled developmental evaluations, and that act as a screen for the presence of such delay across four developmental dimensions. The high predictive ability of CARE (Sensitivity = 95%, Specificity = 85%) in typical children of our sample but at risk of not reaching all their cognitive potential because of social and economic limitations allow considerations for future studies to investigate the measurement of the social skills for the detection of possible early signs of autism spectrum disorder (ASD) in toddlers.

However, further research is necessary to evaluate if limitations related to the sample size and sampling methodology might invalidate these possibilities, such as adding an analysis report on whether the sensitivity and specificity values obtained in CARE vary with children's age. Consequently, the overall results and item analysis of the current study should be interpreted with caution. All suggested diagnostic properties and patterns of agreement and disagreement in the data should be considered exploratory.

Most notably, the final sample and the small within-group numbers demonstrate the effects of demographic variables and item-level results that might be corrected with a large and randomized selected sample. Future research is needed to examine specific skills that are under- or over-reported, and the influence of parents and interviewers' characteristics, like the information on the clinical status of the parents, on the agreement between parent reports and direct testing.

Finally, screening and diagnostics using parent reports as part of long-reach monitoring for social and cognitive developmental status require an examination of engagement and attrition levels of the participants. Previous literature reported parental

engagement by an average completion rate across all cognitive intervention sessions (Haine-Schlagel and Escobar-Walsh, 2015). The average rate is for 49% of participants to abandon the process before cognitive interventions, with a range from 19 to 89%. Haine-Schlagel and Escobar-Walsh's (2015) research indicates that in our case, the 14.9% not returning CARE forms (i.e., attrition) for a non-clinical intervention is very good, but would still reward future inquiry about this issue. Recent studies dedicated to Spanish-monolingual US Latino parents' engagement in an evidence-based program focused on promoting sensitive, responsive parenting for socioeconomically disadvantaged families (So et al., 2020) indicated distinct barriers (e.g., employment challenges, health-related challenges) and facilitators (e.g., knowing other mothers in the group, interest in the program topics), none of which were explored in the current study with CARE.

Further studies should examine whether direct observation at home affects individual development status, and what differences might appear when CARE is not only delivered as a screening tool but structured as an intervention. A comparison with structured interventions will provide a preliminary idea of whether instruments like CARE affect children's outcomes simply by giving caregivers indications to observe and report a broad spectrum of developmental interactions, as do the Guide for Monitoring Child Development (GMCD) and other tools used in global programs (Faruk et al., 2020).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the studies involving human participants were

reviewed and approved by the board at the Faculty of Psychology (Facultad de Psicología) and the General Directorate of Research (Dirección General de Investigaciones) of the Universidad de la Sabana granted ethical approval for the study (Acta CAG #1517 of 19/11/2015). Permission for data collection was granted in agreement with the legal ruling of resolution N° 008430 of 1993 of the Ministerio de Salud de la República de Colombia (Health Ministry of Colombia), which sets out ethical, scientific, technical and administrative norms for research activity with human participants. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

JG-H developed the CARE booklet, supervised the collection and data scoring, performed the statistical analysis, coordinated, and drafted the manuscript. GS participated in the study design and data analytic approach and helped to draft the manuscript. Both authors read and approved the final manuscript.

FUNDING

This work was supported the corresponding author dedication by grants awarded by the internal research fund of the Universidad de la Sabana, Ministerio de Ciencia, Tecnología e Innovación de Colombia (Minciencias), Grant #860, for Doctoral studies support. Opinions, findings, and conclusions from this report are those of the authors and do not necessarily reflect the views of Minciencias.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.725146/full#supplementary-material>

REFERENCES

- Alcantud, F., Alonso, Y., and Rico, D. (2015). Validez y fiabilidad del Sistema de Detección Precoz de los Trastornos del Desarrollo: 3 a 36 meses [Validity and reliability of the early detection system for developmental disorders: 3 to 36 months-old]. *Revista Española de Discapacidad* 3, 107–121. doi: 10.5569/2340-5104.03.01.06
- Bakker, M., and Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behav. Res. Methods* 43, 666–678. doi: 10.3758/s13428-011-0089-5
- Barger, B., Rice, C., Wolf, R., and Roach, A. (2018). Better together: Developmental screening and monitoring best identify children who need early intervention. *Disabil. Health J.* 11, 420–426. doi: 10.1016/j.dhjo.2018.01.002
- Bennetts, S., Mensah, F., Westrupp, E., Hackworth, N., and Reilly, S. (2016). The agreement between parent-reported and directly measured child language and parenting behaviors. *Front. Psychol.* 7:1710. doi: 10.3389/fpsyg.2016.01710
- Boggs, D., Milner, K., Chandna, J., Black, M., Cavallera, V., Dua, T., et al. (2019). Rating early child development outcome measurement tools for routine health programme use. *Arch. Dis. Childhood* 104, S22–S33. doi: 10.1136/archdischild-2018-315431
- Bornstein, M. (2015). "Children's parents," in *Ecological settings and processes in developmental systems*, eds M. H. Bornstein and T. Leventhal (Hoboken, NJ: Wiley Publication), 55–132.
- Britto, P. R., Lye, S. J., Proulx, K., Yousafzai, A. K., Matthews, S. G., Vaivada, T., et al. (2017). Nurturing care: promoting early childhood development. *Lancet* 389, 91–102.
- Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liskowski, U., Behne, T., et al. (2011). Early social cognition in three cultural contexts. *Monogr. Soc. Res. Child Dev.* 76, 1–142. doi: 10.1111/j.1540-5834.2011.00603.x
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* 2, 637–644. doi: 10.1038/s41562-018-0399-z
- Campaña, J., Gimenez-Nadal, J., and Molina, J. (2020). Self-employed and employed mothers in Latin American families: are there differences in paid work, unpaid work, and childcare? *J. Family Eco. Issues* 41, 52–69. doi: 10.1007/s10834-020-09660-5
- Clifton, L., and Clifton, D. A. (2019). The correlation between baseline score and post-intervention score, and its implications for statistical analysis. *Trials* 20:43. doi: 10.1186/s13063-018-3108-3
- Cuartas, J., Rey-Guerra, C., McCoy, D. C., and Hanno, E. (2020). Maternal knowledge, stimulation, and early childhood development in low-income

- families in Colombia. *Infancy Official J. Int. Soc. Infant Stud.* 25, 526–534. doi: 10.1111/inf.12335
- Dawson, P., and Camp, B. W. (2014). Evaluating developmental screening in clinical practice. *SAGE Open Med.* 2:2050312114562579. doi: 10.1177/2050312114562579
- Dias, D. C., Rondon-Melo, S., and Molini-Avejonas, D. R. (2020). Sensitivity and specificity of a low-cost screening protocol for identifying children at risk for language disorders. *Clinics* 75:e1426. doi: 10.6061/clinics/2020/e1426
- Drachler, M., Marshall, T., and de Carvalho-Leite, J. (2007). A continuous scale measure of child development for population-based epidemiological surveys: A preliminary study using item response theory for the denver test. *Paediatr. Perinatal Epidemiol.* 21, 138–153.
- Eng, J. (2014). *ROC analysis: web-based calculator for ROC curves*. Available Online at: <http://www.jrocf.it.org> (accessed September 23, 2019).
- Estrada, E., Ferrer, E., and Pardo, A. (2019). Statistics for evaluating pre-post change: relation between change in the distribution centre and change in the individual scores. *Front. Psychol.* 9:2696. doi: 10.3389/fpsyg.2018.02696
- Faruk, T., King, C., Muhit, M., Islam, M. K., Jahan, I., Baset, K. U., et al. (2020). Screening tools for early identification of children with developmental delay in low- and middle-income countries: a systematic review. *BMJ Open* 10:e038182. doi: 10.1136/bmjopen-2020-038182
- Fernald, L., Prado, E., Kariger, P., and Raikes, A. (2017). *A toolkit for measuring early childhood development in low- and middle-income countries*. Washington, DC: The World Bank.
- Fischer, V., Morris, J., and Martinez, J. (2014). Developmental screening tools: feasibility of use at primary healthcare level in low-and middle-income settings. *J. Health Popul. Nutr.* 32, 314–326.
- Frankenburg, W. K. (1987). *Revised Denver Pre-screening Developmental Questionnaire (PDQII)*. Denver, CO: DDM, Inc.
- Frankenburg, W. K., van Doorninck, W. J., Liddell, T. N., and Dick, N. P. (1976). The denver Prescreening Developmental Questionnaire (PDQ). *Pediatrics* 57, 744–753.
- Fuentes-Biggi, J., Fernandez, I., and Alvarez, E. (1992). *Escala Haizea-Llevant para la evaluación del desarrollo de 0 a 6 años [The Haizea-Llevant scales for the evaluation of development in 0–6 year-olds]*. Vitoria: Gobierno Vasco y Generalitat de Cataluña.
- Gellach, P. (2019). The developmental screening behaviors, skills, facilitators, and constraints of family nurse practitioners in primary care: A qualitative descriptive study. *J. Pediatr. Health Care* 33, 466–477. doi: 10.1016/j.pedhc.2019.01.004
- Gingrich, P. (2004). *Introductory Statistics for the Social Sciences*. Regina, SK: University of Regina.
- Giraldo-Huertas, J., Cano, L., and Pulido, A. (2017). Desarrollo Socio-cognitivo en la primera infancia: los retos por cumplir en Salud Pública en la zona Sabana Centro y Boyacá [Socio-cognitive development in early childhood: the challenges to be met in Public Health in the Sabana Centro and Boyacá area]. *Revista de Salud Pública* 19, 51–57. doi: 10.15446/rsap.v19n4.51787
- Goldfeld, S., and Yousafzai, A. (2018). Monitoring tools for child development: an opportunity for action. *Lancet Global Health* 6, e232–e233. doi: 10.1016/S2214-109X(18)30040-8
- Guan, H., Okely, A. D., Aguilar-Farías, N., Del Pozo Cruz, B., Draper, C. E., El Hamdouchi, A., et al. (2020). Promoting healthy movement behaviours among children during the COVID-19 pandemic. *Lancet Child Adolesc. Health* 4, 416–418. doi: 10.1016/S2352-4642(20)30131-0
- Guevara, J., Gerdes, M., Localio, R., Huang, Y., Pinto-Martin, J., Minkovitz, C., et al. (2013). Effectiveness of developmental screening in an urban setting. *Pediatrics* 131, 30–37.
- Haine-Schlagel, R., and Escobar-Walsh, N. (2015). A review of parent participation engagement in child and family mental health treatment. *Clin. Child Family Psychol. Rev.* 18, 133–150. doi: 10.1007/s10567-015-0182-x
- IBM Corporation. (2017). *IBM SPSS Statistics for Windows, Version 25.0*. Armonk, NY: IBM Corp.
- Ice, A., and Yoldi, M. E. (2002). Psychomotor development of the child and its evaluation in primary care. *Annales Del Sistema Sanitario de Navarra* 25, 35–43.
- Jain, K., Solomon, J., and Ramachandran, S. (2021). Knowledge, attitude and practices on developmental surveillance and screening among health professionals in Indian health care settings: An exploratory sequential mixed methods study. *J. Pediatr. Rehabil. Med.* 14, 55–63. doi: 10.3233/PRM-190649
- Jeong, J., Siyal, S., and Yousafzai, A. K. (2019). Agreement between Fathers' and Mothers' reported stimulation and associations with observed responsive parenting in Pakistan. *Children* 6:114. doi: 10.3390/children6100114
- Johnson, S., Wolke, D., and Marlow, N. (2008). Developmental assessment of preterm infants at 2 years: validity of parent reports. *Dev. Med. Child Neurol.* 50, 58–62.
- Juneja, M., Mohanty, M., Jain, R., and Ramji, S. (2012). Ages and stages questionnaire as a screening tool for developmental delay in Indian children. *Indian Pediatr.* 49, 457–461. doi: 10.1007/s13312-012-0074-9
- Justice, L., Jiang, H., Purtell, K., Schmeer, K., Boone, K., Bates, R., et al. (2019). Conditions of poverty, parent-child interactions, and Toddlers' early language skills in low-income families. *Maternal Child Health J.* 23, 971–978. doi: 10.1007/s10995-018-02726-9
- Khan, N., Muslima, H., Shilpi, A., Begum, D., Akhtar, S., Parveen, M., et al. (2013). Validation of a home-based neurodevelopmental screening tool for under 2-year-old children in Bangladesh. *Child Care Health Dev.* 39, 643–650. doi: 10.1111/j.1365-2214.2012.01393.x
- Kinzler, K., and Spelke, E. (2007). Core systems in human cognition. *Prog. Brain Res.* 164, 257–264.
- Landis, J., and Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.
- Lipkin, P. H., and Gwynn, H. (2007). Improving developmental screening: combining parent and pediatrician opinions with standardized questionnaires. *Pediatrics* 119, 655–657. doi: 10.1542/peds.2006-3529
- Lopez-Boo, F., Cubides-Mateus, M., and Llonch-Sabatés, A. (2020). Initial psychometric properties of the Denver II in a sample from Northeast Brazil. *Infant Behav. Dev.* 58:101391. doi: 10.1016/j.infbeh.2019.101391
- Lu, C., Cuartas, J., Fink, G., McCoy, D., Liu, K., Li, Z., et al. (2020). Inequalities in early childhood care and development in low/middle-income countries: 2010–2018. *BMJ Global Health* 5:e002314. doi: 10.1136/bmjgh-2020-002314
- Ma, S., Song, X., and Huang, J. (2006). Regularized binormal ROC method in disease classification using microarray data. *BMC Bioinform.* 7:253. doi: 10.1186/1471-2105-7-253
- Mahadevan, S., and Broaddus-Shea, E. (2020). How should home-based maternal and child health records be implemented? A global frame- work analysis. *Glob Health Sci. Pract.* 8, 100–113. doi: 10.9745/GHSP-D-19-00340
- McCoy, D. C., Peet, E. D., Ezzati, M., Danaei, G., Black, M. M., Sudfeld, C. R., et al. (2016). Early childhood developmental status in low- and middle-income countries: National, Regional, and Global prevalence estimates using predictive modeling. *PLoS Med.* 13:e1002034. doi: 10.1371/journal.pmed.1002034
- McCoy, D., Waldman, M., Credi Field Team, and Fink, G. (2018). Measuring early childhood development at a global scale: evidence from the caregiver-reported early development instruments. *Early Childhood Res. Q.* 45, 58–68. doi: 10.1016/j.ecresq.2018.05.002
- MESEP-DNP. (2011). *Nueva metodología para la medición de la pobreza monetaria y cifras de pobreza extrema, pobreza y desigualdad 2002-2010. Declaración Comité de Expertos, Declaración de la MESEP*. Available Online at: <http://www.dnp.gov.co/LinkClick.aspx?fileticket=DXInD1TENeU%3d&tabid=337> (accessed June 11, 2016).
- Miller, L. E., Perkins, K. A., Dai, Y. G., and Fein, D. A. (2017). Comparison of parent report and direct assessment of child skills in toddlers. *Res. Autism Spectrum Disord.* 41, 57–65. doi: 10.1016/j.rasd.2017.08.002
- Milner, K. M., Bhopal, S., Black, M., Dua, T., Gladstone, M., Hamadani, J., et al. (2019). Counting outcomes, coverage and quality for early child development programmes. *Arch. Dis. Childhood* 104, S13–S21. doi: 10.1136/archdischild-2018-315430
- Ministerio de Salud de la República de Colombia. (2016). *Actualización y ajuste de la escala abreviada de desarrollo como un instrumento de apoyo en la valoración clínica de desarrollo de los niños menores de siete años. [Updating and adjustment of the abbreviated developmental scale as a support instrument in the clinical assessment of development of children under seven years of age]*. Bogotá: Ministerio de Salud de la República de Colombia.
- Mullen, E. M. (1995). *Mullen Scales of Early Learning*. Circle Pines, MN: American Guidance Service Inc.
- Muñoz Caicedo, A., Zapata-Ossa, H. J., and Pérez-Tenorio, L. M. (2013). Validación de criterio de la Escala Abreviada del Desarrollo (EAD-1) en el dominio audición-lenguaje [Criterion validation of the Abbreviated

- Development Scale (EAD-1) in the hearing-language domain]. *Revista de Salud Pública* 15, 386–397.
- Murphy, R., Jolley, E., Lynch, P., Mankhwazi, M., Mbukwa, J., Bechange, S., et al. (2020). Estimated prevalence of disability and developmental delay among preschool children in rural Malawi: Findings from “Tikule Limodzi,” a cross-sectional survey. *Child Care Health Dev.* 46, 187–194. doi: 10.1111/cch.12741
- Nadeem, S., Avan, B., and Rafique, G. (2016). Development of child assessment and caregiver advice manual for front line health workers to enhance early child development in developing world. *J. Child. Dev. Disord.* 2, 6–15.
- NASEM. (2019). *Monitoring Educational Equity*. Washington, DC: The National Academies Press.
- Ortiz, N. (1991). *Escala Abreviada de Desarrollo [Abbreviated Scale of Development]*. Bogotá: Editorial Ministerio de Salud de Colombia.
- Ozturk-Ertem, I., Krishnamurthy, V., Mulaudzi, M. C., Sguassero, Y., Bilik, B., Srinivasan, R., et al. (2019). Validation of the International Guide for Monitoring Child Development demonstrates good sensitivity and specificity in four diverse countries. *Acta Paediatr.* 108, 1074–1086. doi: 10.1111/apa.14661
- Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford: Oxford University Press.
- Price, P., Jhangiani, R., Chiang, I.-C., Leighton, D., and Cuttler, C. (2017). *Research Methods in Psychology*. Washington, DC: Washington State University.
- Profamilia. (2010). *Encuesta Nacional de Demografía y Salud ENDS. [National Survey of Demographics and Health]*. Available Online at: <http://www.profamilia.org.co/encuestas/Profamilia/Profamilia/> (accessed February 17, 2012).
- Richter, L. M., Daelmans, B., Lombardi, J., Heymann, J., Boo, F. L., Behrman, J. R., et al. (2017). Investing in the foundation of sustainable development: pathways to scale up for early childhood development. *Lancet* 389, 103–118. doi: 10.1016/S0140-6736(16)31698-1
- Richter, L., Cappa, C., Issa, G., Lu, C., Petrowski, N., and Naicker, S. (2020). Data for action on early childhood development. *Lancet* 396, 1784–1786.
- Rivas, S., Sobrino, A., and Peralta, F. (2010). Weaknesses and strengths in assessing early childhood programmes: an assessment of an early childhood Spanish trilingual programme in two- to three-year-old children. *Early Child Dev. Care* 180, 685–701. doi: 10.1080/03004430802231562
- Rubio-Codina, M., and Grantham-McGregor, S. (2020). Predictive validity in middle childhood of short tests of early childhood development used in large scale studies compared to the Bayley-III, the Family Care Indicators, height-for-age, and stunting: A longitudinal study in Bogota, Colombia. *PLoS One* 15:e0231317. doi: 10.1371/journal.pone.0231317
- Ruopp, M., Perkins, N., Whitcomb, B., and Schisterman, E. (2008). Youden Index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biomet. J. Biomet. Zeitschrift* 50, 419–430. doi: 10.1002/bimj.200710415
- Sánchez-Torres, R. (2015). Descomposiciones de los cambios en la pobreza en Colombia 2002-2012 [Decompositions of changes in poverty in Colombia 2002-2012]. *Revista Desarrollo y Sociedad* 75, 349–398. doi: 10.13043/dys.75.9
- Schafer, G., Genesoni, L., Boden, G., Doll, H., Jones, R., Gray, R., et al. (2014). Development and validation of a parent-report measure for detection of cognitive delay in infancy. *Dev. Med. Child Neurol.* 56, 1194–1201. doi: 10.1111/dmcn.12565
- Schiari, V., Simeonsson, R., and Hall, K. (2021). Promoting developmental potential in early childhood: A global framework for health and education. *Int. J. Environ. Res. Public Health* 18:2007. doi: 10.3390/ijerph18042007
- Shah, A. K., Mullainathan, S., and Shafir, E. (2012). Some consequences of having too little. *Science* 338, 682–685.
- Shah, A. K., Shafir, E., and Mullainathan, S. (2015). Scarcity frames value. *Psychol. Sci.* 26, 402–412.
- Shah, A., Mullainathan, S., and Shafir, E. (2018). An exercise in self-replication: Replicating Shah, Mullainathan, and Shafir (2012). *J. Eco. Psychol.* 75:102127. doi: 10.1016/j.joep.2018.12.001
- Sheeran, L., Zhao, L., Buchanan, K., and Xenos, S. (2020). Enablers and barriers to identifying children at risk of developmental delay: A pilot study of Australian maternal and child health services. *Matern. Child Health J.* 25, 967–979. doi: 10.1007/s10995-020-03077-0
- Slemming, W., Cele, R., and Richter, L. (2021). Quality of early childcare in the home and cognitive development at age 5: results from the South African birth to Twenty Plus cohort study. *Early Child Dev. Care* 2021, 1–14. doi: 10.1080/03004430.2020.1868449
- So, M., Almeida Rojo, A. L., Robinson, L. R., Hartwig, S. A., Heggs Lee, A. R., Beasley, L. O., et al. (2020). Parent engagement in an original and culturally adapted evidence-based parenting program, Legacy for Children™. *Infant Mental Health J.* 41, 356–377. doi: 10.1002/imhj.21853
- Sparrow, S., Cicchetti, D., and Balla, D. (2005). *Vineland adaptive behavior scales*. 2. Circle Pines, MN: American Guidance Service.
- Tann, C., Kohli-Lynch, M., Nalugya, R., Sadoo, S., Martin, K., Lassman, R., et al. (2021). Surviving and thriving early intervention for neonatal survivors with developmental disability in Uganda. *Infants Young Child.* 34, 17–32. doi: 10.1097/IYC.0000000000000182
- Trude, A., Richter, L., Behrman, J., Stein, A., Menezes, A., and Black, M. (2021). Effects of responsive caregiving and learning opportunities during pre-school ages on the association of early adversities and adolescent human capital: an analysis of birth cohorts in two middle-income countries. *Lancet Child Adolesc. Health* 5, 37–46. doi: 10.1016/S2352-4642(20)30309-6
- Van Keer, I., Bodner, N., Ceulemans, E., Van Leeuwen, K., and Maes, B. (2020). Parental behavior and child interactive engagement: a longitudinal study on children with a significant cognitive and motor developmental delay. *Res. Dev. Disabil.* 103:103672. doi: 10.1016/j.ridd.2020.103672
- Vanbelle, S. (2017). Comparing dependent kappa coefficients obtained on multilevel data. *Biomet. J. Biomet. Zeitschrift* 59, 1016–1034. doi: 10.1002/bimj.201600093
- Villagomez, A. N., Muñoz, F. M., Peterson, R. L., Colbert, A. M., Gladstone, M., MacDonald, B., et al. (2019). Neurodevelopmental delay: Case definition & guidelines for data collection, analysis, and presentation of immunization safety data. *Vaccine* 37, 7623–7641. doi: 10.1016/j.vaccine.2019.05.027
- Vitrikas, K., Savard, D., and Bucay, M. (2017). Developmental delay: When and how to screen. *Am. Family Phys.* 96, 36–43.
- WHO. (2020). *WHO guideline: improving early childhood development, 2020*. Available Online at: <https://www.who.int/publications/i/item/improvingearlychildhooddevelopment-whoguideline> (accessed March 13, 2020)
- Zanca, F., Hillis, S. L., Claus, F., Van Ongeval, C., Celis, V., Provoost, V., et al. (2012). Correlation of free-response and receiver-operating-characteristic area-under-the-curve estimates: results from independently conducted FROC/ROC studies in mammography. *Med. Phys.* 39, 5917–5929. doi: 10.1118/1.4747262

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Giraldo-Huertas and Schafer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Advances in Behavioral Remote Data Collection in the Home Setting: Assessing the Mother-Infant Relationship and Infant's Adaptive Behavior via Virtual Visits

Eunkyung Shin¹, Cynthia L. Smith² and Brittany R. Howell^{1,2*}

¹ Fralin Biomedical Research Institute at Virginia Tech Carilion, Roanoke, VA, United States, ² Department of Human Development and Family Science, Virginia Tech, Blacksburg, VA, United States

OPEN ACCESS

Edited by:

Sho Tsuji,
The University of Tokyo, Japan

Reviewed by:

Marion I. van den Heuvel,
Tilburg University, Netherlands
Holly Rayson,
Centre National de la Recherche
Scientifique (CNRS), France

*Correspondence:

Brittany R. Howell
brhowell@vtc.vt.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 April 2021

Accepted: 09 September 2021

Published: 01 October 2021

Citation:

Shin E, Smith CL and Howell BR
(2021) Advances in Behavioral
Remote Data Collection in the Home
Setting: Assessing the Mother-Infant
Relationship and Infant's Adaptive
Behavior via Virtual Visits.
Front. Psychol. 12:703822.
doi: 10.3389/fpsyg.2021.703822

Psychological science is struggling with moving forward in the midst of the COVID-19 pandemic, especially due to the halting of behavioral data collection in the laboratory. Safety barriers to assessing psychological behavior in person increased the need for remote data collection in natural settings. In response to these challenges, researchers, including our team, have utilized this time to advance remote behavioral methodology. In this article, we provide an overview of our group's strategies for remote data collection methodology and examples from our research in collecting behavioral data in the context of psychological functioning. Then, we describe the design and development of our strategies for remote data collection of mother-infant interactions, with the goal being to assess maternal sensitivity and intrusiveness, as well as infants' adaptive behaviors in several developmental domains. During these virtual visits over Zoom, mother-infant dyads watched a book-reading video and were asked to participate in peek-a-boo, toy play, and toy removal tasks. After the behavioral tasks, a semi-structured interview (Vineland Adaptive Behavior Scale – VABS III) was conducted to assess the infant's adaptive behavior in communication, socialization, daily living skills, and motor domains. We delineate the specific strategies we applied to integrate laboratory tasks and a semi-structured interview into remote data collection in home settings with mothers and infants. We also elaborate on issues encountered during remote data collection and how we resolved these challenges. Lastly, to inform protocols for future remote data collection, we address considerations and recommendations, as well as benefits and future directions for behavioral researchers in developmental psychology research.

Keywords: remote data collection, behavioral observation, home setting, infant adaptive behavior, mother-infant relationship

INTRODUCTION

During the COVID-19 pandemic, investigators have faced challenges in conducting research, with traditional face-to-face data collection methods having been paused or otherwise disrupted. Social distancing mandates and safety barriers forced researchers to shift in person data collection in laboratories to remote data collection in other settings (Sy et al., 2020). Thus, observational

measures were restricted during the COVID-19 pandemic. While this restriction caused many disruptions to traditional behavioral assessment data collection, this unique situation also forced researchers to consider novel research designs and to develop remote data collection methodology.

Despite challenges in data collection during this time of social distancing, advances in technology, such as the increased access to synchronous web-based video conferencing platforms (e.g., Zoom and Skype), have allowed for innovative ways of collecting behavioral data that may compensate for the lack of, or extend, traditional face-to-face data collection methods. Even prior to restrictions on in person data collection, remote data collection methods have been implemented in behavioral research (Strickland et al., 2003). However, most studies have predominantly focused on qualitative research including online interviews and focus groups (Archibald et al., 2019). Few reports have been published about observational data collection using web-based video conferencing platforms in naturalistic settings. The purpose of the current report is to share strategies and experiences in remote data collection in naturalistic settings using video conferencing platforms to enrich the methods available to collect behavioral data during this global health crisis. In this article, we delineate the specific strategies that we applied to integrate laboratory tasks and a semi-structured interview into remote data collection in home settings with mothers and infants. We also elaborate on issues encountered during remote data collection and how we resolved these challenges. Lastly, we address considerations and recommendations for behavioral researchers in child development research to inform optimized protocols for future remote data collection.

Web-Based Data Collection

Remote data collection provides greater flexibility and effectiveness in time, cost, and access to participants. In qualitative research on participants' perception about research using Skype (Lo Iacono et al., 2016), participants mentioned that they prefer taking part in research at home to traveling to the laboratory in terms of the amount of time that they spend for research. Participants can save time and the cost of traveling to the laboratory, and researchers can also have flexibility in timing and space where they conduct research. In particular, given that mother-infant interaction is more likely to be related to infant's feeding and sleeping schedules, and/or child temperament characteristics that may make being in unfamiliar spaces stressful (Graag et al., 2012), comfortable space and thoughtful scheduling are necessary to accurately capture mother-infant interactions in daily life. Resolving these logistical issues allows researchers to access geographically diverse and disadvantaged populations, as long as researchers accommodate access to internet tools and environment (Sy et al., 2020). Thus, remote data collection can be used for rural populations and cross-cultural studies with better access to participants who face challenges to in-person participation, including reduced mobility or large geographical spread.

Two types of web-based data collection technology have been utilized in the past – asynchronous and synchronous (Berg, 2007). Asynchronous methods support web-based

communication at different times such as email or online surveys. Scott and Schulz (2017) developed an asynchronous online platform called Lookit, to collect infants' preferential looking paradigms. Parents participated in self-administered tasks with their children at their convenience by accessing the Lookit website without live interaction with researchers. Lookit is available for researchers to conduct their own research via Github Projects¹. Recently, Rhodes et al. (2020) conducted unmoderated remote research in which parents and their children participated in online software using families' webcams without involvement of researchers. Items about gender stereotypes and parent-child conversations about gender were conducted using the online software. A study setting without direct interaction with researchers putatively elicits more natural behavior from families because of the absence of strangers (Rhodes et al., 2020). Resources used for implementation of the study have been shared on the following website².

In contrast, synchronous methods include real-time interactions such as online messengers and video conference calls (e.g., Zoom and Skype) that enable back-and-forth exchange of interactions (Sullivan, 2012). Sheskin and Keil (2018) developed a video chat platform to validate the method by replicating standard developmental tasks with children aged between 5 and 10 years old. Most children in their study presented correct answers in social tasks and causal reasoning tasks. Because synchronous methods transmit verbal and non-verbal cues through real time video and audio, researchers are better able to replicate the features of face-to-face in-person interactions using these technologies. We implemented a synchronous method because live interaction with researchers allows the researcher to conduct the study in a consistent way across all participants (Sheskin and Keil, 2018). Although few studies using online data collection have been published, there have been efforts to advance the field of online assessment. For example, "Many Babies-At Home"³ is a methodological project in which researchers developed cross-cultural online testing of infants. In this project, multiple laboratories across the world have collaborated to develop and distribute universal and robust practices in online testing methods for developmental studies.

Current Study

In this article, we describe the application of remote data collection in a natural home setting through a video conferencing platform to share our experiences with researchers who are considering new remote data collection methods. We provide an overview of our group's strategies for remote data collection methodology and examples from our research in collecting behavioral data in the context of psychological functioning. Then, we describe the design and development of our strategies for remote data collection of mother-infant interactions, with the goal being to assess maternal sensitivity and intrusiveness, as well as infants' adaptive behaviors in several developmental domains. We delineate the specific strategies we applied to integrate

¹github.com/orgs/lookit/projects/

²discoveriesonline.org

³<https://manybabies.github.io/MB-AtHome>

laboratory tasks and a semi-structured interview into remote data collection in home settings with mothers and infants. We also elaborate on issues encountered during remote data collection and how we resolved these challenges. Lastly, to inform protocols for future remote data collection, we address considerations and recommendations, as well as benefits and future directions for behavioral researchers in developmental psychology research.

IMPLEMENTATION OF REMOTE DATA COLLECTION

We originally planned to invite mothers and their infants to our lab to conduct a 10-min free play session in which mothers and infants interact with a standardized set of toys. Additionally, the Bayley Scales of Infant and Toddler Development – Fourth edition (Bayley-4; Bayley and Aylward, 2019) and episodes from the Laboratory Temperament Assessment Battery (Lab-TAB; Goldsmith et al., 1993) were planned to assess general development, temperament, and mother-infant behavioral interactions. After the outbreak of COVID-19, we had to shift our plan for face-to-face data collection in our laboratory to remote data collection. Given that synchronous web-based video conferencing platforms can capture real-time interaction with a private recording function, we decided to collect behavioral data through a video conferencing platform, specifically Zoom⁴. The remote data collection for our study has been an alternative way of collecting behavioral data in natural setting. In the following sections, we describe the process of how we prepare for and conduct virtual visits to share our experiences in the application of remote data collection in natural settings by utilizing the advantages of technology described previously.

As part of a longitudinal study exploring maternal biobehavioral influence on infant brain and behavioral development, we have been conducting remote virtual visits to collect behavioral observation data on mother-infant interaction and infant adaptive behavior through a Health Insurance Portability and Accountability Act (HIPAA) compliant Zoom platform. We selected Zoom as our video conferencing platform because it offers secure recording and data storage features. Zoom provides real-time encryption of meetings and backup recordings while complying with HIPAA regulations. All procedures were optimized for participants to join using their phones to avoid having to exclude anyone due to not having access to a home computer with a camera and microphone. Mothers were instructed to install the Zoom app on their phone to conduct the virtual visit. In the longitudinal study design, we plan to collect behavioral data when the infants are 3, 6, 18, and 24 months old. We have been conducting remote data collection for the 3- and 6-month-old visits and will conduct in-person data collection for the 18- and 24-month-old visits because laboratory cameras that can adjust angles are better to capture the movement of older infants. For this article, we discuss our approaches to conducting virtual visits with mothers and their 3- and 6-month-old infants; however, tasks and instructions for 6 months are the same

except for the specific book and toy provided. We selected age-appropriate books and toys for each time point.

Preparation for the Virtual Visit

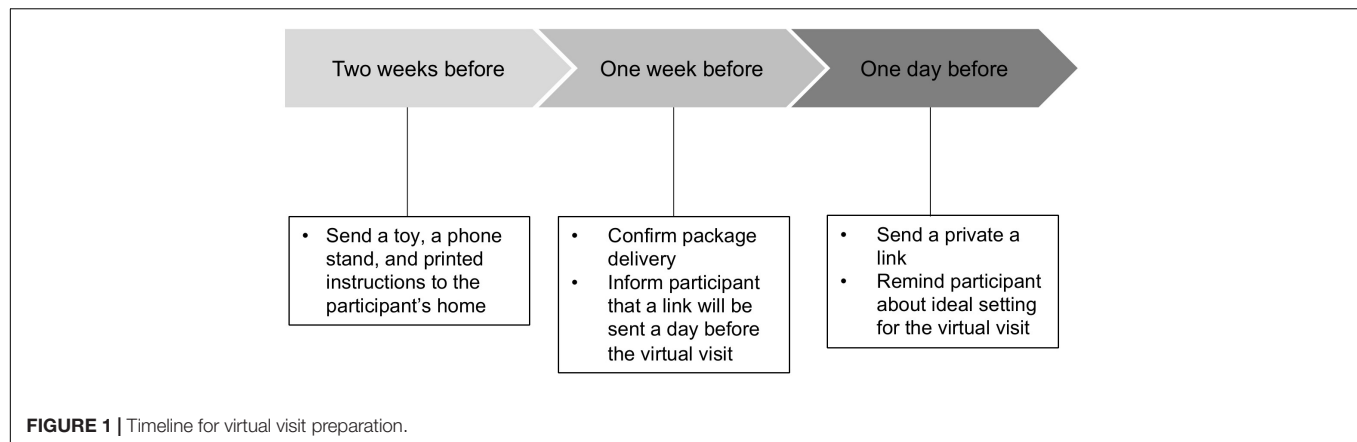
Participants were recruited from previous research and participant referral. During the consent call, a trained research assistant explained the general description of tasks and that their interactions will be recorded. Participants were informed that recorded videos will be stored in the institutional data center with restricted permissions and access. Moreover, the research assistant explicitly told the participants that they could withdraw from the study at any time for any reason and it would not be held against them. After the initial consent call, specific preparations for the virtual visit occurred at three points: 2 weeks, 1 week, and 1 day before the virtual visit (**Figure 1**). Two weeks before the visit, we mailed a toy, a phone stand, and printed instructions to the participant's home. The toy (**Figure 2**) was chosen because it is developmentally appropriate and entertaining, would elicit interaction between mothers and infants, families were unlikely to already have it, and it was easy to mail (i.e., it was available on Amazon Prime). In the instructions (see **Supplementary Appendix 1**), directions for Zoom app installation and example photos of recommended positions the mother and infant should use during the visit were included (**Figure 3**). Mothers were asked to not show the toy that was sent to their baby until the visit to ensure it remained novel to the infant. One week before the visit, we contacted mothers again to confirm that they received the package, and we informed them that a private Zoom link would be sent a day before the visit. A HIPAA-compliant Zoom link was texted to the mother's phone a day before the visit. We also reminded her that the session would be recorded and what the ideal setting for the visit would look like (e.g., a quiet and bright place for the visit, ensuring their phone is fully charged, etc.). Because 3-month-old infants have difficulty sitting unsupported, we asked mothers to use a supportive pillow, a Bumbo seat, or a bouncer during the virtual visit.

During the Virtual Visit

Virtual visits were scheduled with consideration for the infant's feeding and sleep schedule to ensure that the infant was alert and ready to play during the session. Just before the visit, the experimenters turned off their computer notifications and phones and opened the materials for the visit. Once the session began, but before recording began, the mother was informed that the session was going to be recorded and asked to turn on the "do not disturb" mode on her phone. Recording began once she confirmed that she was comfortable with it. She was then asked to position her phone horizontally using the phone stand provided. After positioning her phone as asked, she was asked to troubleshoot a camera angle and position that would capture both her and her infant's faces.

Two research assistants were required for each visit: one research assistant (the experimenter) ran the tasks while the other research assistant (the recorder) recorded the visit. Two experimenters were necessary because during a Zoom meeting in which a person is sharing their screen (in this case, the

⁴<https://zoom.com>



experimenter), the person sharing their screen is unable to adjust the size of the window showing the participant's video (**Figure 4**). The recorder was responsible for maximizing the size of the window displaying the participant's video during the session (**Figure 5**). During the visit, the two research assistants worked together to complete the three tasks including the book reading video, peek-a-boo game, and toy play and removal to assess mother-infant interaction. The experimenter provided a general description of the tasks followed by specific instructions before each task. In addition, the Vineland Adaptive Behavior Scales –

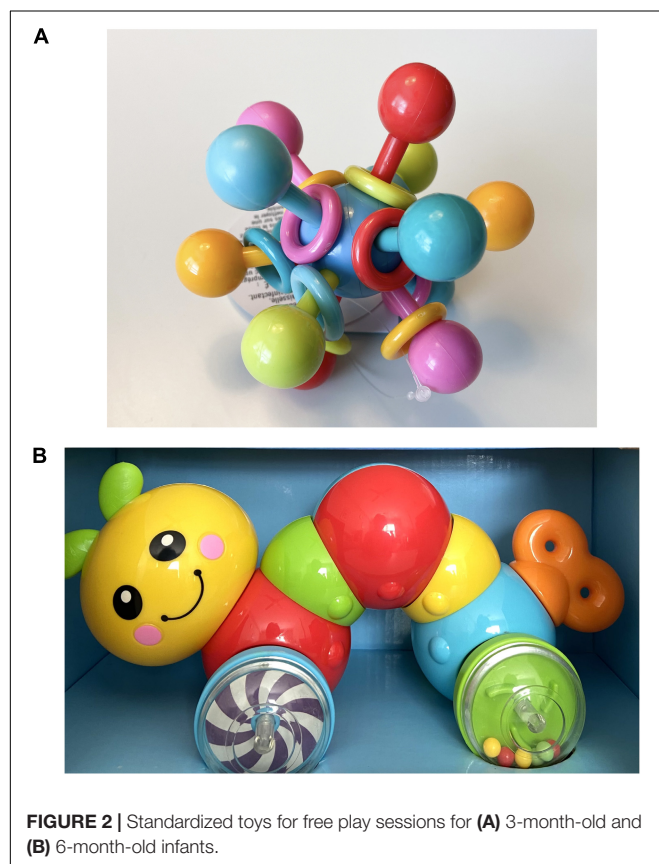
Third Edition (VABS-III; Sparrow et al., 2016), a semi structured parent interview was designed to measure the infant's adaptive behavior (see **Figure 6** for the visit schedule). If at any time the infant became fussy or needed a break, the experimenter allowed a break to calm the infant down. The pace of the visit was determined by the mother-infant dyad's needs and conditions.

Book Reading Video

Mothers and their infants were asked to watch a 2-min video of the book “*Happy Baby, Sad Baby*” by Leslie Patricelli, being read aloud by a female research team member (see **Supplementary Appendix 2**). In the video participants could not see the face of the reader, but could see her hands turning the pages as she read. We chose to have an experimenter read the book instead of mothers to ensure that all infants received the same stimulation related to the book, allowing for a standardized measure of attention (i.e., time spent attending to the book vs. looking away from the book). The mother was asked to sit with her infant in her lap while they watched the video (**Figure 7**). Mothers were instructed not to redirect their infant's attention should they turn away or otherwise stop attending to the video to allow for robust and accurate quantification of infant attention directed toward the video. The experimenter then shared their screen to show the book reading video and turned off their own camera. After the video was over, the mother was prompted to talk to her infant about the book for 1 min. Recordings were labeled using a study ID number free of personally identifiable information. Coding of the interaction will take place at a later time and will include assessments of maternal behavior (i.e., maternal sensitivity and responsiveness) and infant attentiveness (e.g., time spent attending to the video). To account for the confounding effects of screen exposure on infant behavior, we also asked mothers about their infants' exposure to screen media at the end of the visit.

Peek-a-Boo Game

The experimenter then asked the mother to initiate and participate in a 1-min play session of peek-a-boo with her infant (**Figure 8**). At the 3- and 6-month visits, the dyad sat on the floor with the infant supported with pillows or in a highchair if available (**Figure 3** for recommended positions). Once the



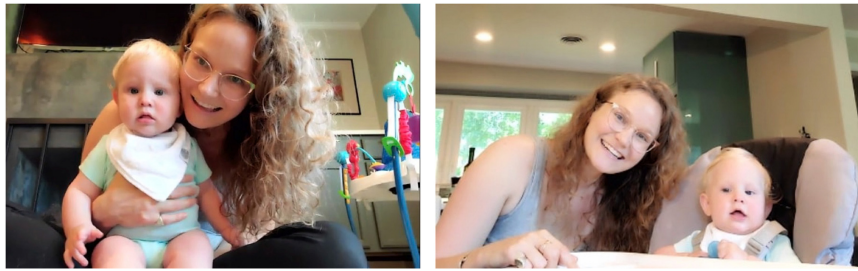


FIGURE 3 | Recommended positions for virtual visit.



FIGURE 4 | Issue with size of participant's screen in Zoom.

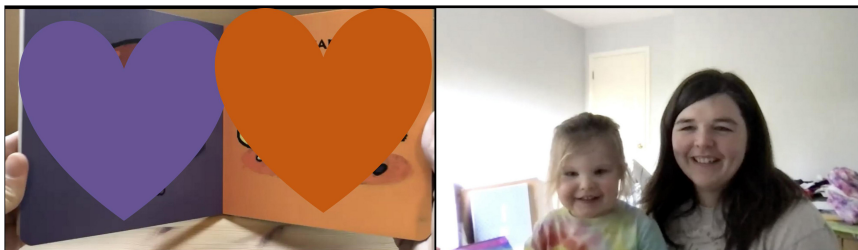


FIGURE 5 | Adjusting size of participant's screen in Zoom.

mother began the peek-a-boo game, the experimenter turned off their camera and muted their audio. This interaction will also be coded at a later time for infant positive and negative affect and maternal sensitivity and intrusiveness.

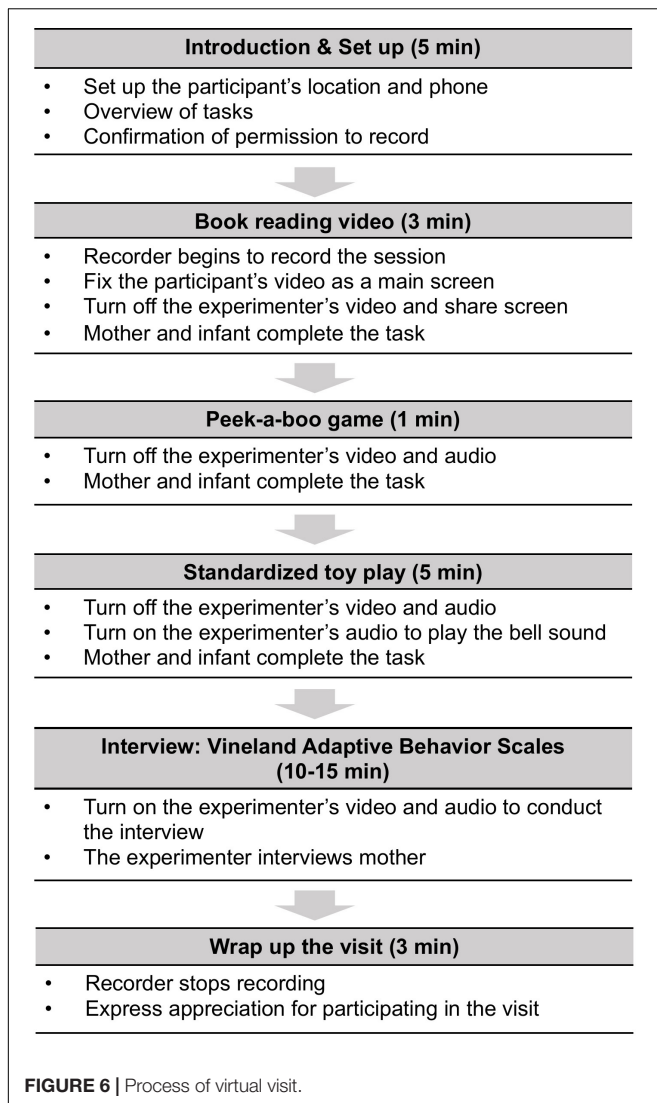
Standardized Toy Play

Mothers and their infants participated in a 2-min play session using the toy that was sent earlier, followed by toy removal for 1 min, and another 2-min play session. Each mother was asked to interact with her infant as she normally would (**Figure 9**). Once the mother began to play with her infant, the experimenter turned off their camera and muted their audio, and a 1-min timer was started. Following the toy play session, the mother was instructed to remove the toy from her infant, but to leave it where her infant

could see, but not reach it, for 1 min. After 1 min, the mother was prompted to give the toy back to her infant and to play for two additional minutes. Infant positive and negative affect and maternal sensitivity and intrusiveness will be coded at a later time.

Infant's Adaptive Behavior

The Vineland Adaptive Behavior Scales – Third Edition (VABS-III; Sparrow et al., 2016) assessed child adaptive behavior in several developmental domains: communication, socialization, daily living skills, and motor. The VABS-III was designed to be administered by an experimenter as a semi-structured interview with a caregiver, in this case the mother. The interview usually takes 10–15 min to conduct. Because the infant portions of the session were completed, infants could



stay or go to another caregiver during the interview. The VABS-III consists of behaviors that infants display without physical help or reminders. Both the experimenter and the recorder were trained to administer this measure, and both rated the mother's answers using a printed questionnaire. We used Pearson's web-based system (Q-global) for test administration and scoring. Q-global supports both management of examinees' records and production of specific and comprehensive reports of automatically calculated scores. After the experimenter entered their scores for each item in Q-global, the recorder verified the scores that the experimenter entered and published a total score report.

CHALLENGES AND RECOMMENDATIONS

In this section, we describe difficulties that we encountered, and recommendations and considerations that facilitated

remote data collection in the home setting using video conferencing tools. First, researchers need to confirm that the participant's technological environment is sufficient for remote data collection. Because synchronous video conferencing tools use the internet; therefore, stable internet connectivity and quality are the first requirements for both the research team and participants to collect data remotely using the procedures described in this manuscript. Most video conferencing platforms used on phones work an average internet speed between 60 and 100 kbps (McNally, 2020). In addition, other devices and phone capabilities including microphones and cameras should be checked before the visit. Researchers can check the participant's technological environment during the consent session, or can schedule a separate practice session in the same setting as the visit. If the participant's technological environment is insufficient for remote data collection, the research team can offer technological aids. For example, if the quality of the internet connection is insufficient, the research team can lend the participant a Wi-Fi hotspot or other devices to meet technological requirements. In addition, there are many public libraries that loan Wi-Fi hotspots to community members.

Second, an instruction document or checklist of logistical set up requirements helps participants prepare for the virtual visit. Unlike the laboratory setting, researchers are less able to control the space where participants' behavior is observed. We ask participants to find a quiet and uninterrupted place for the visit, charge their phone, and turn on "do not disturb" mode on their phones. A simple instruction booklet including pictures facilitated participants' set up of logistical requirements (see Instructions in **Supplementary Material**). In particular, example pictures of good camera angles and providing a phone stand with markings of specific angles aided participants in following recommended settings (see **Figure 3**). Moreover, a short tutorial video may facilitate standardized participation. For example, the Emerging Minds Lab at Arizona State University shared a tutorial video for a cognitive task via Twitter⁵.

Third, as mentioned earlier, an effective way to ensure the virtual setting will work for the visit is by running practice sessions. Because device malfunction, video or audio issues, and lack of Zoom experience can disrupt remote data collection, practice sessions prepare a research team to set up web-based remote data collection. We have had several mock virtual visits among research team members and two pilot sessions with mother-infant dyads before we started to collect remote data. It was through a pilot session that we discovered the screen proportion issue described previously when the experimenter shared their screen. In the book reading task, we needed to ensure the participant's screen was large enough to code their behavior. However, once the experimenter shared their screen while recording the session, the shared screen was bigger than the participant's screen, even though we pinned the participant's screen (**Figure 4**). We consulted with the "Many Babies-At Home (MBAH)" group, and a researcher from MBAH proposed an applicable solution to address this issue. To adjust the participant's screen size, another research assistant (the recorder)

⁵<https://twitter.com/EmergingMindsAZ/status/1281618737079017473?s=20>

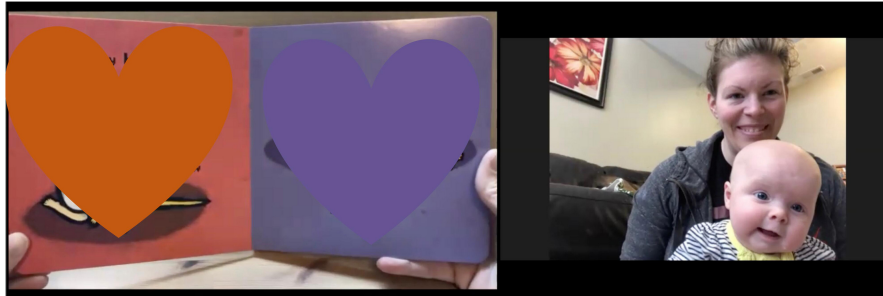


FIGURE 7 | Data from a book reading video.

joined the session to record the visit. Since the recorder did not share their screen, the recorder was able to make the participants' face as large as possible (**Figure 5**). Because remote data collection is currently underutilized as an observational data tool, sharing

challenges with other researchers is a productive way to resolve issues and advance methodological skills.

Fourth, researchers should assure that video conferencing platforms offer privacy and secure service for confidentiality of data. Due to significant increases in video conference meetings during the COVID-19 pandemic, uninvited outsiders have more opportunities to enter meetings and interrupt the session (e.g., Zoom bombing). To protect the session and participants, meeting access should be protected by a password, or a research team can use features which control the attendee's entrance and terminate meeting sessions (i.e., the waiting room feature). Moreover, HIPAA established privacy and security standards must be maintained to protect personal privacy. Researchers need to ensure that the video conferencing platforms utilized provide HIPAA-compliant services. For example, the HIPAA-compliant version of Zoom uses safeguards to prevent any unauthorized access in their environment to meet these HIPAA requirements. Other than privacy and secure service for confidentiality of data, the principles of ethical issues in online data collection are similar to in-person contexts. Lobe et al. (2020) mentioned that "researchers who already have approval their review board will probably only need to file a simple amendment to their original proposal to shift from in-person to online data collection" (p. 5). Common ways to obtain participant's consent for remote data collection are consent phone call or conference call and email consent form to participants. Research teams ask for scanned signatures or use electronic signature programs such as DocuSign⁶ to collect participants' signatures. For example, we explained study protocols and answer questions that participants had during a consent call and send a link to a consent form which participant can then sign. During a consent call, our participants were informed about the recording of the sessions, the private and secure data storage, and their right to withdraw from the study at any time for any reason without penalty.

DISCUSSION

Prior to social distancing guidelines, which led to challenges for inviting participants into the lab, we planned to observe mother-infant interactions and assess infants' development



FIGURE 8 | Data from a peek-a-boo game.



FIGURE 9 | Data from a standardized toy play.

⁶<https://www.docusign.com>

in person. However, in the COVID-19 era, we needed to find alternative ways to pursue answers to important and pressing research questions. Video conferencing platforms are able to concurrently record back-and-forth exchanges of interactions in a private internet setting. Additionally, time and location flexibility allowed us to consider mothers and infants' schedules at home. We were able to adjust our observational measures to include synchronous behavioral assessments and a semi-structured interview with mothers to collect infants' developmental information without any attrition so far. Visual inspection suggests that the quality of data obtained through the virtual visit has been similar to data obtained from the lab setting; mother and infants faces are clearly visible, allowing for consistent coding, and the same study materials were used across participants (see **Figures 7–9**). When participants veered away from prescribed camera angles or protocols, an experimenter guided mothers to conduct the study in a consistent way across all participants. A link to an example video of a 3-month virtual visit is provided in **Supplementary Appendix 3**. As we experienced, shifting from traditional face-to-face data collection to remote data collection required careful consideration of conceptual and logistical aspects of data collection. In this report, we describe the application of remote data collection in a natural home setting through a video conferencing platform to share our experiences with researchers who are considering new remote data collection methods.

Because remote data collection through video conferencing platforms is still a nascent topic, there are limitations and careful considerations for future research. For example, web-based platforms require digital tools and knowledge and internet connectivity. These requirements might overlook populations that lack access to technology tools or confidence in using them. Even though the digital environment has been rapidly developing, it is important to consider underrepresented groups who struggle with technology to gain generalizable knowledge. In a qualitative study using Zoom (Archibald et al., 2019), most participants encountered some challenges with joining the session. Researchers need to support the use of technology with approachable instructions and tools, as demonstrated here. Another limitation might be that families may not want to participate from their homes for a variety of reasons. Because home environments reflect families' lifestyles, there may be participants who do not want to share this view into their homes. In this case, researchers can suggest other places, such as public libraries, for families to participate. Although it may vary depending on the library, most libraries offer private rooms for community members to reserve.

Here we only focused on using the HIPAA-compliant version of the Zoom platform. Practical features and considerations could be different depending on videoconferencing platforms. It is important to consider the functions that will best convey a project's needs and institutional support. There are several options for remote data collection. For example, Webex⁷ has also been widely used for research and Skype⁸ is common for

interpersonal communication. Researchers could also utilize the HD video feature through GoToMeeting⁹. Lobe et al. (2020) proposed the following criteria be considered when choosing a videoconferencing platform: "the number of participants in a same session, audio/video recording, one-click access for participants, and privacy features (p. 3)." Researchers also need to find secure data storage in accordance with ethical procedures. Prior literature has recommended that recording data through the internet should be saved in local storage (i.e., the researcher's computer) and not the cloud storage provided by the platform to preserve the third party's privacy (Lobe et al., 2020).

Despite limitations, remote data collection through videoconferencing platforms offers opportunities for researchers to pursue data collection until social distancing recommendations are relaxed, and beyond. Significant increases in access to electronic devices and the internet across the world, improvements in the platforms, and sharing practical guidelines among researchers promise to advance the effective use of remote data collection. Researchers can increase rigor by utilizing advantages of technologies, detailed and approachable instructions with careful considerations, practice sessions, and electronic safeguards. It is our hope that sharing our experiences and issues in remote data collection with mothers and infants with other researchers will extend methodological tools to historically underrepresented populations (i.e., rural populations) in developmental science.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Virginia Polytechnic Institute and State University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s) for the publication of any identifiable images or data presented in this article.

AUTHOR CONTRIBUTIONS

ES: protocol design, data collection, study execution, and writing – original draft preparation. CS: protocol design and writing – reviewing and editing. BH: protocol design, writing – reviewing and editing, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

⁷ www.webex.com

⁸ www.skype.com

⁹ www.gotomeeting.com

FUNDING

BH is an iTHRIV Scholar. The iTHRIV Scholars Program is supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Numbers UL1TR003015 and KL2TR003016.

ACKNOWLEDGMENTS

We thank the members of the Howell Lab at the Fralin Biomedical Research Institute at Virginia Tech Carilion who contributed to virtual visits including Jamie Holt, Collin Gregg, Ahmad Obaidi, as well as Eliza Joy Howell, Alex DiFeliceantonio,

Isaac DiFelice Howe, Jade Brooks, Maebry Brooks, Colleen Smith, and Nora Smith who contributed to article figures. We also thank MinJu Kim at the University of California, San Diego who provided advice on our data recording Zoom issue. We are indebted to the families who so generously participated in our research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.703822/full#supplementary-material>

REFERENCES

- Archibald, M. M., Ambagtsheer, R. C., Casey, M. G., and Lawless, M. (2019). Using zoom videoconferencing for qualitative data collection: perceptions and experiences of researchers and participants. *Internat. J. Q. Methods* 18, 1–8. doi: 10.1177/1609406919874596
- Bayley, N., and Aylward, G. (2019). *Bayley Scales of Infant and Toddler Development*, 4th Edn. San Antonio, TX: Harcourt Assessment, Inc.
- Berg, B. L. (2007). *Qualitative research methods for the social sciences*, 6th Edn. Boston, MA: Pearson's Education, Inc.
- Goldsmith, H. H., Reilly, J., Lemery, K. S., Longley, S., and Prescott, A. (1993). *Preschool Laboratory Temperament Assessment Battery (PS Lab-TAB; Version 1.0). Technical Report, Department of Psychology, University of Wisconsin—Madison*.
- Graag, J. A., Cox, R. F., Hasselman, F., Jansen, J., and de Weerth, C. (2012). Functioning within a relationship: Mother–infant synchrony and infant sleep. *Infant Behav. Dev.* 35, 252–263. doi: 10.1016/j.infbeh.2011.12.006
- Lo Iacono, V., Symonds, P., and Brown, D. H. (2016). Skype as a tool for qualitative research interviews. *Sociolog. Res. Online* 21, 103–117. doi: 10.5153/sro.3952
- Lobe, B., Morgan, D., and Hoffman, K. A. (2020). Qualitative data collection in an era of social distancing. *Internat. J. Q. Methods* 19, 1–8. doi: 10.1177/1609406920937875
- McNally, C. (2020). *What internet speed do I need for Zoom?* Available online at: <https://www.reviews.org/internet-service/zoom-technical-requirements/> (accessed date 2020, December 1).
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Scott, K., and Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Sheskin, M., and Keil, F. (2018). TheChildLab.com A Video Chat Platform for Developmental Research. *Charlottesville, VA: Society for the Improvement of Psychological Science*. doi: 10.31234/osf.io/rn7w5
- Sparrow, S. S., Cicchetti, D. V., and Saulnier, C. A. (2016). *Vineland adaptive behavior scales*, 3rd Edn. London: Pearson.
- Strickland, O. L., Moloney, M. F., Dietrich, A. S., Myerburg, S., Cotsonis, G. A., and Johnson, R. V. (2003). Measurement issues related to data collection on the World Wide Web. *Adv. Nurs. Sci.* 26, 246–256.
- Sullivan, J. R. (2012). Skype: An appropriate method of data collection for qualitative interviews? *Hilltop Rev.* 6, 54–60.
- Sy, M., O'Leary, N., Nagraj, S., El-Awaisi, A., O'Carroll, V., and Xyrichis, A. (2020). Doing interprofessional research in the COVID-19 era: a discussion paper. *J. Interprof. Care* 34, 600–606. doi: 10.1080/13561820.2020.1791808

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Shin, Smith and Howell. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Designing Virtual, Moderated Studies of Early Childhood Development

Liesbeth Gijbels^{1,2†}, Ruofan Cai^{1,2†}, Patrick M. Donnelly^{1,2} and Patricia K. Kuhl^{1,2*}

¹Department of Speech & Hearing Sciences, University of Washington, Seattle, WA, United States, ²Institute for Learning & Brain Sciences, University of Washington, Seattle, WA, United States

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, United States

Reviewed by:

Jenny Saffran,
University of Wisconsin-Madison,
United States
Keith Apfelbaum,
The University of Iowa, United States

*Correspondence:

Patricia K. Kuhl
pkkuhl@uw.edu

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 12 July 2021

Accepted: 06 September 2021

Published: 11 October 2021

Citation:

Gijbels L, Cai R, Donnelly PM and
Kuhl PK (2021) Designing Virtual,
Moderated Studies of Early
Childhood Development.
Front. Psychol. 12:740290.
doi: 10.3389/fpsyg.2021.740290

With increased public access to the Internet and digital tools, web-based research has gained prevalence over the past decades. However, digital adaptations for developmental research involving children have received relatively little attention. In 2020, as the COVID-19 pandemic led to reduced social contact, causing many developmental university research laboratories to close, the scientific community began to investigate online research methods that would allow continued work. Limited resources and documentation of factors that are essential for developmental research (e.g., caregiver involvement, informed assent, controlling environmental distractions at home for children) make the transition from in-person to online research especially difficult for developmental scientists. Recognizing this, we aim to contribute to the field by describing three separate moderated virtual behavioral assessments in children ranging from 4 to 13 years of age that were highly successful. The three studies encompass speech production, speech perception, and reading fluency. However, varied the domains we chose, the different age groups targeted by each study and different methodological approaches, the success of our virtual adaptations shared certain commonalities with regard to how to achieve informed consent, how to plan parental involvement, how to design studies that attract and hold children's attention and valid data collection procedures. Our combined work suggests principles for future facilitation of online developmental work. Considerations derived from these studies can serve as documented points of departure that inform and encourage additional virtual adaptations in this field.

Keywords: virtual, moderated, development, early childhood, online

INTRODUCTION

Over the past decades, technological advancements have expanded the scale and scope of academic research. A body of literature between 1995 and 2005 proposed a series of benefits and disadvantages associated with the initial wave of Internet-based research (Hewson et al., 1996; Reips, 2001, 2002; Duffy, 2002; Kraut et al., 2004), which underscored a time when digital research was relatively novel and small-scale. Despite the growing popularity of much online work following the rise of digital media in the 21st century, research in the field of child development stayed relatively resistant, and digital formats of developmental research have only recently been demonstrated (Scott et al., 2017; Scott and Schulz, 2017; Sheskin and Keil, 2018; Gweon et al., 2020; Nussenbaum et al., 2020; Rhodes et al., 2020; Sheskin et al., 2020).

Further, established methodological adaptations in this field are largely characterized as immature, especially in the adoption and validation of online behavioral assessments (Scott and Schulz, 2017; Nussenbaum et al., 2020; Rhodes et al., 2020).

In 2020, as the COVID-19 pandemic led to reduced social contact, causing many research laboratories to close, the scientific community began to investigate online research methods that would allow continued work. Remote, digital modalities have been recognized as viable substitutions for in-person research settings (Reips, 2001, 2002; Sheskin and Keil, 2018). In comparison with laboratory-based research methods, advantages associated with general online research (e.g., reduced operating costs, increased access to diverse populations, and reduction in experimenter effects) have been reported (Reips, 2002; Bohner et al., 2002). Accompanying the recent rising trend of remote research practice, these advantages make it possible to envision a future of advanced remote methodologies for developmental work.

However, shifting from in-person to remote modalities is not without challenges. For example, Reips (2002) identified experimental control and attrition as common concerns in online research. In particular, remote behavioral measures tend to introduce additional confounds which are often attributed to increased variability in research environment and equipment. Further, online adaptations of developmental studies require nuanced, age-specific considerations such as accounting for children's attention span and cognitive load in the task design and administration (Gibson and Twycross, 2008).

Although solutions have been proposed to address some of the challenges (Reips, 2002), peer-reviewed methodological reports of adaptation from in-person to online developmental studies are rather limited, awaiting substantial input. Recognizing the lack of documented observations from existing virtual research and its potential to deter future implementations of online developmental work, we aim to contribute to the field by describing three researcher-moderated virtual assessments in children ranging from 4 to 13 years of age, encompassing assessments of their speech processing skills and reading fluency. The varied domains, in combination with the age groups targeted by each study, required different methodological approaches. However, the success of our remote adaptations shared certain commonalities regarding informed consent, study designs that attract and hold children's attention, and valid data collection procedures. Through this work, we hope to suggest principles for future facilitation of online developmental research, and we believe that considerations derived from these three studies can serve as documented points of departure that inform and encourage additional virtual adaptations in this field.

The three studies included in this paper sought to adapt their original in-person task designs for remote facilitation with researcher moderation. While the moderated format was appropriate for these studies, both moderated and unmoderated designs have their pros and cons, and we encourage developmental scientists to make decisions with regard to the degree of moderation while facilitating online child studies. Compared to moderated studies, unmoderated or fully automated studies are less work-intensive during the research appointments, but

it may require more preparation work in task automation and involve additional steps of data processing. Elimination (or lessening) of researcher involvement is advantageous in bias removal, as it is often replaced by consistent machine-delivered instructions. This facilitates the comparison across replications of unmoderated studies (Rhodes et al., 2020). However, for the same reason that makes unmoderated formats appealing to some, the lack of researcher real-time involvement also presents several challenges.

Informed Consent and Data Security

Ethics of non-therapeutic research involving children are a delicate issue, as children are vulnerable and would likely not benefit directly from participation (Lambert and Glacken, 2011). In language suitable for the intended individual, informed consent/assent should communicate the study's purpose and procedures, associated benefits and risks, confidentiality, safety, etc. Additionally, when appropriate, the researcher or caregiver may need to verbally communicate the informed consent, which is often crucial to ensuring participants' understanding, as the informed consent ought to be viewed as a process rather than a product, beyond signature collection (Whitehead, 2007; Gibson and Twycross, 2008).

For many virtual studies, using online applications, such as REDCap, are an appealing way to collect e-consent and to build and manage online databases. A lot of web tools come with built-in privacy measures, allowing digital consent to be completed efficiently and stored securely. On platforms such as Pavlovio and Gorilla, documentation of major identifying information can stay detached from research data, and it is often possible to record the consent process and data collection separately (Sheskin and Keil, 2018). However, it is generally difficult for unmoderated consent processes to create space for researchers to interact with participants and address participants' questions or concerns. In addition, experimental processes that rely on human-machine interactions (e.g., text-based or video/audio recording) alone could run a higher risk of technical error, resulting in corrupted recordings, for example. In contrast, a moderated process enables candid researcher-participant communication and provides flexibility for procedural adjustments (guided by a well-designed rubric), which is frequently needed due to increased variability and unpredictability of virtual studies in home environments.

Protecting participants' privacy and data confidentiality is among the top priorities in human subject research. Remote consent processes in recent years have shown varying formats. Some researchers opt for digital acquisition of text-based consent *via* email (Nussenbaum et al., 2020) or online secure databases (Donnelly et al., 2020a, 2020b), and others acquire verbal consent and assent using automated video and audio recording (Scott and Schulz, 2017; Rhodes et al., 2020). While research moderation is not required for either option, the latter, when unmoderated, is subjected to technical issues with video/audio recording, potentially resulting in invalid data if not detected promptly (Rhodes et al., 2020). Scott and Schulz (2017) reported that up to 16% of their data were discarded due to inadequate

consent recordings. In contrast, in addition to audio or video recording documentations (Sheskin and Keil, 2018), researcher observation and natural dialogues during moderated consent procedures help the researcher detect and address technical issues and ensure understanding of informed consent.

In addition, experimental stimuli and research data that are delivered and collected digitally are subjected to additional ethical scrutiny, specifically regarding data security. Some study designs may require transportation of research equipment or digital transfer of data files. In these cases, encrypting the devices and data files (e.g., using passwords or proprietary software) can significantly lower security risks, and related considerations are growing in prominence as new technologies increasingly deliver utility in research methods. As our capabilities are being enhanced rapidly, the scientific community needs to continually assess the implications of technologically enabled advancements in human subject research.

Experimental Control and Parental Involvement

Additionally, experimental control concerns are presented in traditional research settings and highlighted even more in virtual environments. For example, whereas it is fairly straightforward to manipulate the acoustic environment in a laboratory's sound booth, it is impossible to obtain the same level of control in participants' homes. A realistic attempt would be to instruct caregivers to prepare a "quiet room" for the research appointment. In addition to audible noises, families may have different levels of visual and tactile distractions at home (e.g., siblings or pets). Furthermore, unless experimental equipment is specified or provided for the participants, technical device differences (e.g., headphones, Internet connection stability, screen sizes) also need to be considered.

Motivation and Sustained Attention

Probably one of the main reasons for the slow move to online research in developmental work is that experimental designs involving children are typically more complex than those involving adults. A major challenge for child development researchers is how to best engage participants, remove distractions, and motivate participation given age-specific attention spans.

Interactions between the participant and researcher may be helpful in maintaining the child's interest level. Developmental research studies, especially ones targeting auditory or visual perception, can benefit from researcher observation even if the task itself is fully automated. In a moderated session, the researcher-observer would be able to note any circumstances or issues that might come up and adjust as needed, whether it be troubleshooting technical difficulties, regulating caregiver involvement, clarifying task instructions, or introducing necessary breaks.

Adapting developmental research for online environments inevitably introduces tangible changes to a study's experimental design and setup, but perhaps equally important is its impact on a socio-psychological aspect of human subject research,

the researcher-participant relationship. Traditionally in a laboratory environment, face-to-face interactions can often motivate participation. While social interactions through a screen are often perceived as "flattened" and cannot fully replace their in-person counterparts, it is still possible to enhance researcher-participant relationships and to foster participant engagement and motivation through researcher moderation of remote studies. Notably, in studies involving children who struggle with unfamiliar surroundings (e.g., children with autism), the introduction of a stranger (i.e., the researcher) and a new environment (i.e., the laboratory) can be intimidating at times and interfere with the validity in data collection. In these cases, virtual assessment is an especially advantageous alternative, as it allows for in-home research participation, and can reduce or remove the perception of stranger interaction (Rhodes et al., 2020).

Validity of Online Adaptations

Given the variety of developmental behavioral work and the limited resources for online adaptations available, questions arise regarding the validity of these adaptations. Several attempts have been made to compare in-person and remote work (Sheskin and Keil, 2018; Rhodes et al., 2020; Yeatman et al., 2021), which highlighted some important questions. Considerations for task design, stimulus presentation, attention maintenance, and results interpretation are all crucial to ensure a study's validity. A good example that warrants caution is the interpretation of norm-referenced tasks when assessed remotely. Examples of these are intelligence tests, reading assessments, vocabulary assessments, etc. Although big companies like Pearson assessments have started to offer some tasks remotely with written guidelines, they warrant against interpretation of the norms:

A spectrum of options is available for administering this assessment *via* telepractice; however, it is important to consider the fact that the normative data were collected *via* face-to-face assessment. Telepractice is a deviation from the standardized administration, and the methods and approaches to administering it *via* telepractice should be supported by research and practice guidelines when appropriate (Pearson, 2021).

As such, interpretation of these norms when moved online should be deliberated prior to implementation.

In this paper, three different virtual studies will be discussed. Each study was initially conceived and developed for in-person environments and subsequently moved online. The original laboratory-based research plans will be summarized, along with adaptations made to enable remote facilitation. The studies targeted different questions and distinct age groups, which led to different approaches. Although the results of these studies are very promising and will each contribute to their field independently, the focal point of this paper is the adaptations we made to the three studies (Section "Procedural Modifications for Online Studies"), our data regarding their success and validity (Section "Methods; Developing Remote-Friendly Measures for Moderated, Developmental Studies"), and our resulting perspective on future implementations of virtual studies

(Section “Discussion”). Through this paper, our ultimate aim is to motivate a continuance of remote developmental research, post-pandemic.

PROCEDURAL MODIFICATIONS FOR ONLINE STUDIES

To represent the vast array of developmental research in this paper, we selected three distinct studies that varied in research goals and participants’ demographics. An Imitation study (see Section “Assessment of Vocal Imitation of Native and Nonnative Vowels (Cai and Kuhl, in Prep.)”) focusing on speech acquisition (age 4), an Audiovisual (AV) study (see Section “Audiovisual Speech Processing in Relationship to Phonological and Vocabulary Skills Gijbels et al., in Press.”) focusing on speech perception (age 6–7), and a Reading study (see Section “A Symbolic Annotation of Vowel Sounds for Emerging Readers (Donnelly et al., 2020b)”) focusing on bringing digital tools completely online (age 8–13) will be described. Each study’s research questions, study designs, and modifications made for their virtual implementation will be outlined in Section “Procedural Modifications for Online Studies”, specific methodological adaptations will be expanded further in Section “Methods: Developing Remote-Friendly Measures for Moderated, Developmental Studies”, and the three studies will be joined together in Section “Discussion” to draw general guiding principles for future online behavioral research.

Assessment of Vocal Imitation of Native and Nonnative Vowels (Cai and Kuhl, in Prep.)

Vast differences have been observed in second language (L2) learners’ ability to imitate novel sounds – while the majority of learners exhibit and maintain a foreign accent throughout their lifetime, some are able to produce accurate L2 pronunciations to a near-native level. These individual differences have been previously characterized as largely innate and fixed (Abrahamsson and Hyltenstam, 2008). While a number of recent published accounts have attempted to identify, in part, correlates of this talent variability (Christiner and Reiterer, 2013; Hu et al., 2013; Franken et al., 2015; Ghazi-Saidi and Ansaldo, 2017), efforts have been somewhat scattered. And despite its prevalence to the foundational research in speech perception and production, vocal imitation remains an understudied topic.

In this study, we investigated four-year-old typically developing (TD) children’s ($N=57$) ability to imitate vowel sounds, both native and nonnative, to understand young children’s sensorimotor knowledge of speech. The intent of the study was to understand how children’s ability to imitate speech relates to age, language history, and other environmental factors. The specific aims were to: (1) measure the acoustic details of children’s imitated vowels and assess the acoustic distance between their productions and those of the model they were imitating (2) determine whether children’s abilities differed for

native vs. nonnative vowels, and (3) investigate individual differences in speech imitation ability among young children.

A laboratory-based format of this study was carried out during the initial pilot phase. Upon arrival at the laboratory, parents were first asked to complete a questionnaire, which surveyed environmental factors such as socio-economic status and language background. Then, the speech imitation task involving child participants was administered *via* an animal puppet theater set up in a sound booth. To deliver the auditory and visual stimuli, the researcher operated the animal puppet’s mouth behind the puppet theater, “lip-syncing” the puppet to pre-recorded speech sounds played through the speakers. A research assistant sat beside the child facing the puppet theater and assisted the participant as needed. Two video cameras, a pair of audio speakers, and a studio-quality microphone were set up in the booth. In an observation room next door, caregivers were invited to watch the live task procedures on a TV screen. This setup allowed parents to stay informed of their children’s behaviors or needs while avoiding unnecessary interference to the study session. This procedure worked during the pilot stage of this experiment, and 4-year-old children demonstrated their ability and willingness to engage in the task.

In response to the public health crisis posed by COVID-19, the study was adapted digitally to accommodate remote testing. We modified the parental survey format, the protocol for parental involvement, and the means of video and audio recording of experimental sessions. Parental questionnaires were conducted digitally using a secure online portal, and the speech imitation task took place over Zoom. In the modified, online version of the imitation task, instead of plush puppets, participants interacted with animal cartoon characters on the researcher’s computer screen (*via* screen share), repeating vowel sounds after them, some “native,” and some “nonnative” to the child’s language (see **Figure 1**). In speech perception and production experiments, developing reliable audio systems is central to achieving consistent stimulus presentation and quality data acquisition. The key measurement in this study is the acoustic distance between the vowel target (i.e., model) and imitation (i.e., production), which is calculated using formant frequency values of the vowel target and of the imitation. Recognizing the variability in hardware and software configurations across participants, in addition to using the video and audio recording system built into the Zoom video conferencing platform, we also mailed individual pocket-sized audio recorders – the Language ENvironment Analysis system (LENA™, the LENA Research Foundation, Boulder, CO) – to the participating families to capture the children’s speech productions in their environment more accurately and consistently. Additionally, given the participants’ young age and the virtual administration of an interactive task, parents assisted with facilitation of the appointment when needed.

Online adaptations of the study were successfully implemented. Forty-six out of 57 participating subjects were included in the analysis, with a resulting total of over 7,000 utterances examined, and audio files retrieved from the LENA recorders provided adequate acoustic information for the purpose of vowel formant analysis (see “Validity of online adaptations”).

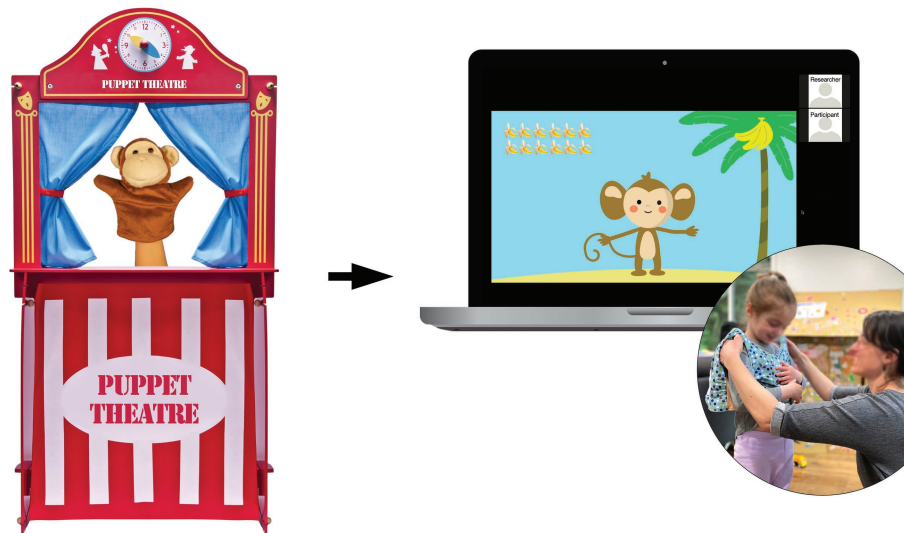


FIGURE 1 | Visualization of the Imitation Study. Digital cartoon animation adapted from in-lab puppet theater setup, used to deliver auditory stimuli remotely via Zoom during the imitation task. Speech data collected via LENA vests and recorders worn by child participants (Cai and Kuhl, in prep).

Audiovisual Speech Processing in Relationship to Phonological and Vocabulary Skills

The benefits of audiovisual (AV) speech perception, more specifically, having access to the (Gijbels et al., in press) articulation movements when the auditory speech signal is degraded by noise, have been well studied in adults (see Grant and Bernstein, 2019 for a review). And although we know that infants (Kuhl and Meltzoff, 1984) and children (Lalonde and Werner, 2021 for a review) are sensitive to AV speech information, the size and the presence of an actual AV speech benefit have been debated (Jerger et al., 2009, 2014; Fort et al., 2012; Ross et al., 2011; Lalonde and McCreery, 2020). More specifically, 5- to 8-year-olds show highly variable results when completing audiovisual speech perception tasks. As suggested by Lalonde and Werner (2021), these results might be explained by extrinsic factors as task complexity, intrinsic factors (i.e., individual developmental skills) or the combination of both (i.e., general psychophysical testing performance).

The specific aims of this study were to assess (1) whether TD children in first grade ($N=37$; 6–7 years old) show AV speech enhancement in a noisy environment when a task is presented with low cognitive and linguistic demands (i.e., extrinsic factors) (2) whether individual variability in AV gain is related to intrinsic developmental factors (Jerger et al., 2009; Ross et al., 2011), or to (3) the combination of intrinsic and extrinsic factors. To address these questions, the participants completed an AV speech perception task (see Figure 2). In this task, audio-only (i.e., stimulus word + speech-weighted noise + still image), audiovisual (i.e., stimulus word + speech-weighted noise + matching video), or visual-only (i.e., speech-weighted noise + video) stimuli were presented in 200 trials, broken up in 10 blocks. The stimulus was followed by four answer options (i.e., one a correct answer, two options were

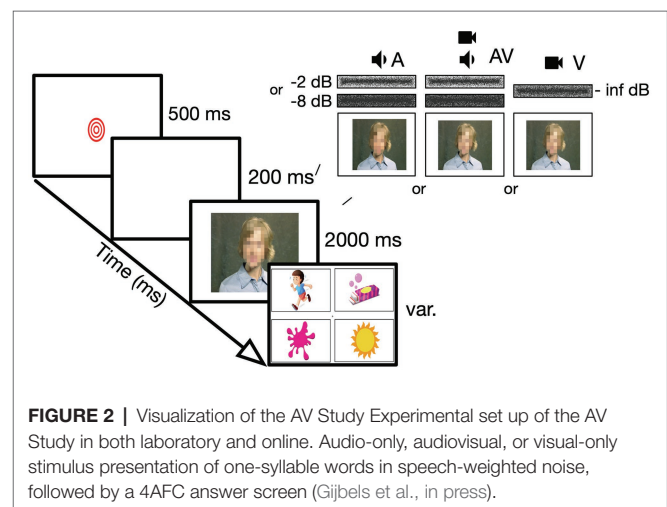


FIGURE 2 | Visualization of the AV Study Experimental set up of the AV Study in both laboratory and online. Audio-only, audiovisual, or visual-only stimulus presentation of one-syllable words in speech-weighted noise, followed by a 4AFC answer screen (Gijbels et al., in press).

related in word form, and a random answer option). Additionally, participants completed standardized measures of vocabulary (Expressive Vocabulary Test; EVT-3; Williams, 2019) and phonological awareness skills (Phonological and Print Awareness Scale; PPA; Williams, 2014), and a third control auditory psychophysical task, that was very similar in setup to the AV task but had no speech or visual component to it. The cognitive and linguistic demands were limited by using a closed set (four-alternative forced choice; 4AFC) picture pointing task, with a stimulus set of consonant-vowel-consonant words that are well known by typically developing children of this age (Holt et al., 2011).

In a laboratory setting, we would measure individuals' behavioral and psychophysical performance in a quiet, controlled environment (i.e., sound booth) to ensure the reliability of stimuli and response. Conducting the tasks in

a quiet room in the laboratory provides the opportunity to assess baseline control of hearing thresholds and visual acuity, eliminates potential interference (e.g., background noise), avoids unintended asynchrony of auditory and visual stimuli, and maintains exact output levels and quality of all stimuli using a calibrated computer. It also allows interpretation of normed behavioral tests, as they can be assessed according to the manual. Interference from parents would be limited as they would wait in the waiting room and instructions and assessment would be provided by a trained research assistant.

For both the in-person and the online version of the experiment, the stimulus presentation followed by 4AFC answer options would look identical. Also, the number of breaks (stimulus blocks) and catch trials were kept consistent. However, to move the tasks to a virtual environment, the tools for stimulus presentation (i.e., assessment format), data interpretation methods, and parental involvement had to be re-envisioned. Participants would complete the tasks at home, in front of their personal computer in a varied environment (i.e., background noise). Parents were instructed before and during the moderated session to provide a “controlled” and consistent environment. They would act as technical support and report presented technical hiccups, but also take over tasks that the research assistant would normally provide in the laboratory (e.g., providing mouse control when the child had insufficient computer handiness). Parents would provide information about hearing and vision of the participant *via* an online parental questionnaire, rather than collecting this “objectively” in-person. The psychophysical tasks would now be collected directly *via* an experiment builder (i.e., Lab.js; Henninger et al., 2020). This provided the quality of stimulus presentation that had a close resemblance to in-person testing. The disadvantage of working directly in the experiment builder was that parents had to download the results and email it to the researcher. The experiment builder allowed us to have consistent and pre-recorded instructions and pre-assembled stimuli that assured the simultaneous presentation of audio and video (as discussed in 3.4). Since it was not possible to control the exact output level of the stimuli on the participants’ computer, we provided an opportunity for participants to set their individual computer to a level that was comfortable and kept consistent throughout the tasks. The main aim of this study was to see whether children this age showed AV enhancement. This was determined by subtracting participants’ overall percentage correct score of the audio-only trials (i.e., speech in noise combined with a still image) from the percentage correct score from all AV trials (i.e., speech in noise combined with a matching video of the woman speaking). Therefore, results were interpreted as relative levels (i.e., difference in percentages), rather than absolute hearing thresholds. The behavioral tasks, that is, vocabulary (EVT) and phonological awareness (PPA), were assessed using similar methods to in-person testing, over Zoom. Raw scores were used, rather than normed standardized scores due to the limited knowledge of norm interpretation in an

online setting (as discussed in 3.7.2). Attention control (catch trials¹ and random answer options in the 4AFC task) were built-in. Although this is always important when working with children, we focused a bit more on the importance of attentional control online. We note that there are very little data about attentional behavior for online tasks with children (as discussed in 3.6).

A Symbolic Annotation of Vowel Sounds for Emerging Readers

Although there is an extensive market for educational technologies for literacy (Guernsey and Levine, 2015; Donnelly et al., 2020b), the vast majority of these technologies lack an evidenced-based component (Guernsey and Levine, 2015; Christ et al., 2018) and show small effect sizes (Cheung and Slavin, 2011). It is too often assumed that new implemented technologies will simply be successful. Meta-analyses reveal, however, limited short- and long-term gains, small sample sizes, and less rigorous designs (Blok et al., 2002; Stetter and Hughes, 2010; Grant et al., 2012). Despite this too often assumed “digital magic,” we can specify how technology advantages emerging readers by examining its many opportunities for practice, feedback, motivation, and autonomy in the learning process (Soe et al., 2000; Richardson and Lyytinen, 2014; Wolf et al., 2014; Ronimus and Lyytinen, 2015; Benton et al., 2018; McTigue et al., 2020). And more interestingly, technology provides a platform to supplement more classical learning with individualized materials that struggling readers require, both inside and outside of the classroom. This leads to empowering shared experiences with caregivers.

This study investigated the efficacy of an educational technology to support literacy in 8- to 13-year-old struggling readers ($N=78$), as characterized by performance on a battery of reading assessments. The technology used was specifically designed to scaffold and empower emerging readers (at home and in school). *Sound it Out* is a web-based educational application focusing on phonological awareness and letter-sound correspondence skill. It utilizes visual cues to vowel identity that are placed under the words to scaffold grapheme-phoneme correspondence during connected text reading and was studied in a randomized controlled trial design. As seen in **Figure 3**, the tool provides visual cues for all vowels in a given text: for example, under the “ou” in “you,” the image of a moon is provided to cue the sound /u/ in /mun/. The aims of the study were to determine whether extended practice with visual cues could produce measurable gains in reading skill. More specifically (1) Can a digital annotation inspired by evidence-based reading practice help children decode novel words, and (2) can this tool help children read more fluently? Lastly, we were interested whether (3) children’s gains were impacted by supervised practice with a caregiver. The study began in the laboratory, with participants asked to attend three in-person appointments for assessment/training with two two-week practice

¹The catch trials were created by showing a presentation of the cartoon character on top of the stimulus video or image. The children had to yell the cartoon’s name, and this was noted by researchers and parents

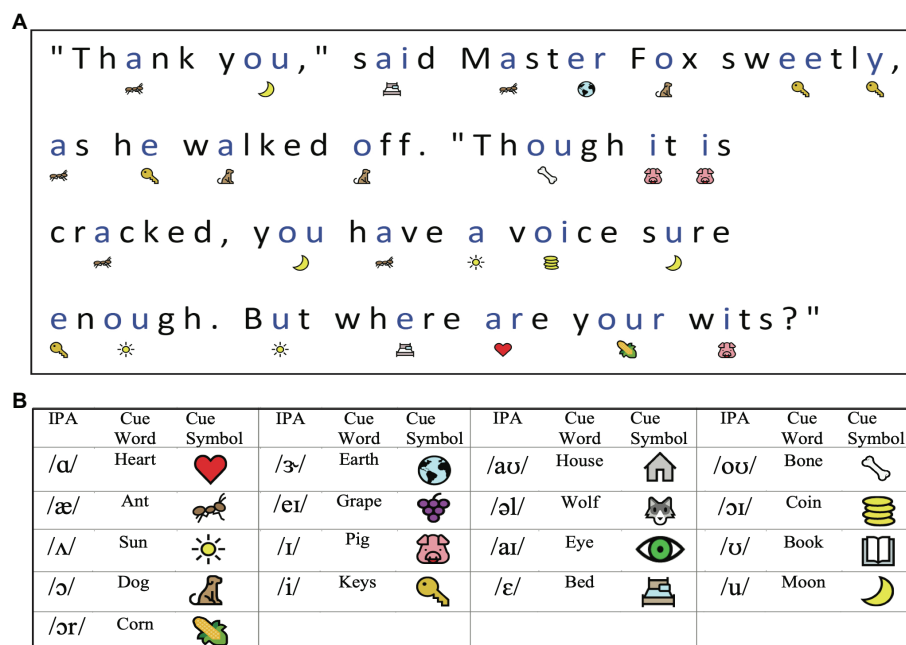


FIGURE 3 | Visualization of the Reading Study: *Sound it Out* provides visual cues under each word that prompts readers on the pronunciation of the vowels contained in the words. Panel (A) presents a sample of a fable passage with symbolic annotations. Panel (B) shows the legend of the image cues used. (Donnelly et al., 2020b).

periods at home in between. With the onset of the COVID-19 pandemic, the study was moved online, and this affected study logistics as well as data collection and training fidelity.

The digital literacy app studied was aimed at supporting phonological decoding for both isolated word reading and connected text fluency. In the laboratory research setting, instruction for both child participants and caregivers occurred in-person, with shared attention to teaching materials and a blend of digital/hardcopy materials to maximize learning. Moreover, assessment involved the use of a standard device (tablet) that reduced variability and controlled for potential issues of screen size, resolution, font size, and Internet connectivity. During the first session, all participants (3 groups) completed baseline tests in an uncued condition (without the *Sound it Out* tool). The two intervention groups would then receive training on the app (for more detail see Donnelly et al., 2020a), one group with active caregiver involvement, and one without. These groups would do instructed at home training and come back to the laboratory for a retest session (session 2), by using the cued condition of the app. A refresher training was provided, and another 2 weeks with training were repeated to end in a final session 3. The control group completed an identical trajectory, without the cues in the app and without caregiver involvement. We collected five outcome measures at all three time points: decoding accuracy, real-word decoding, pseudo-word decoding, passage reading accuracy, and passage reading rate.

By moving to a virtual setting, the methodology was amended with impacts to the training program, the approach to assessment,

and investments in device distribution. Where we could provide the same tablet for all participants in the laboratory, we now offered children the use of their own tablet if preferred. Additionally, all tests were presented digitally, where they were on paper for the in-person version. This added some extra measures to ensure digital consistency, visual presentation of reading passages, and test materials. These adjustments extended the online visits, with a prolonged start to ensure adequate assessment. Another time-intensive aspect was moving the training instructions online. Where initially, the child (and parent) would share a view of the tablet with the researcher who guided them through the app, visually and verbally, they now had to be guided verbally *via* a second screen (a computer) with videoconferencing. Training instruction could be provided at an equal level (as discussed in 3.7.3), but it was definitely more time-intensive.

METHODS; DEVELOPING REMOTE-FRIENDLY MEASURES FOR MODERATED, DEVELOPMENTAL STUDIES

In order to align with remote research modalities, critical adjustments were made to each study (See **Appendix**, Table A). For example, all three studies required changes to their respective informed consent procedures and operating logistics. Moreover, in individual task procedures, adjustments were made

to presentation mode, video/audio recording format, behavioral measures, and attention maintenance.

Informed Consent and Privacy

To ensure participants' understanding, in-person consent procedures are commonly guided by researchers, providing time and space to emphasize or clarify information on the informed consent, such as affirming the participant's right to withdraw from the study at any time, as well as to address questions and concerns from participants. Comparable procedures can be carried out in virtual studies. Video conferencing platforms (e.g., Zoom, Microsoft Teams, Google Meet) have brought well-appreciated convenience in enabling researchers to moderate consent procedures and online tasks. However, certain privacy and security issues have also been exposed amid the soaring popularity of these platforms. While such issues are heavily dependent upon the individual software's safety protocol, much responsibility in protecting research subjects lies within institutions and researchers. In our three studies, the research appointments were conducted over Zoom, and for online security purposes, we generated and assigned passcodes and an online waiting room, and to start off the appointments, we reviewed our video/audio protocol with the participant's caregiver to ensure comprehension of informed consent.

In the Imitation and Reading studies, the majority of the caregivers signed the consent forms prior to the behavioral assessments. For those who were unable to, time was allocated at the beginning of the sessions to address questions and complete consent procedures. In the AV study, parental consent and child assent were both collected *via* audio recording. During all three studies' consent procedures, no identifiable information was collected. Instead, the research teams generated unique aliases (e.g., multi-digit numeric codes, code names, login credentials) for parents to input for anonymous identification. Links were established between the aliases and participant identities, which were only stored on local computers. Additionally, in the AV and Reading studies (where participants were mature enough to understand study procedures and provide meaningful assent), verbal assent was acquired *via* video or audio recordings.

Caregiver Involvement

For child studies in laboratory settings, caregiver involvement is often minimized. During the in-person pilot phase of the Imitation study, parents of the 4-year-old participants were invited to view the experimental process from an observation room. In the original designs of the AV and Reading studies involving older children, parents would be asked to stay in a neighboring waiting room or sit at a distance in the experimental room while the study is in session. These strategies removed possible confounds related to caregiver involvement during the task and allowed parents of younger children to monitor the task process and to attend to the children's needs. When moving these studies online, the caregiver was advised to stay with or near the child during the appointments. Additionally, caregiver

roles varied by participants' age. Among younger children, parental physical assistance is often necessitated for task completion. For instance, to enhance participant compliance, it is typically recommended for a toddler to sit on the parent's lap or beside the parent in front of the computer, whereas older children tend to have sufficient self-control to perform tasks with less caregiver involvement.

Specifying the role of caregivers in our studies was not only critical to ensuring proper consent, privacy, and children's comfort, it also helps control parental involvement across families. As such, it was crucial for caregivers to be briefed on research procedures prior to the appointment. In order to uncover the role of caregiver-supervised practice, the Reading study implemented two training/practice conditions: unsupervised, independent reading and supervised, dyadic reading with a caregiver. In the online implementation, this involved providing consistent instructions for caregivers both during laboratory visits and at-home practice sessions. It was also important for caregivers to know what *not* to do. For instance, in a screening task involving picture naming in the Imitation study, parents were allowed to provide hints when children did not recognize an image but were instructed to avoid using word form (i.e., morphological) variations of the targeted word. Parental assistance is also crucial when a research task requires complex manipulation of digital devices. In the AV and Reading studies, participants had to actively interact with a computer or tablet. Although most children at the age of 8–13 (Donnelly et al., 2020b) were able to perform the required manipulations once the app was set up by the caregiver, children aged 6–7 (Gijbels et al., in press) were not equally skillful in manipulating the mouse/trackpad. Therefore, during a training phase, based on participants' computer proficiency, the researcher made decisions regarding the assistance provided by the parent. If necessary, parents would make mouse clicks, with the limitation that the child had to indicate the answers (by pointing) and the mouse would return to a neutral position in the middle of the screen after every trial.

Typically, parental feedback and parent-guided responses are discouraged in child studies. However, the challenge caused by the unpredictability of caregiver involvement in remote environments can be blunted by deciding prior to the experimental data collection whether parents would assist the child. Because our studies were moderated, researchers could make observations of participants and parents, and as required, instructing parents regarding their participation. In addition to parents receiving instructions at the beginning of each appointment, built-in training phases (as in the Imitation and AV studies) allowed instructions to be repeated to ensure adherence. In addition to the detailed protocols that were verbally communicated to parents prior to the appointments, the research team of the Imitation study also mailed a hardcopy flowchart to help visualize the task procedures. Lastly, because parents are often tempted to help their child "succeed" when they struggle with a task, as the more complex items occur in certain trials, reminder instructions regarding parental intervention were presented throughout the tasks as well.

An additional concern raised with parents is the timing of online appointments. Because they take place in participants' homes, scheduling has to factor in families' daily routines and the degree to which it is possible to participate without interruptions. When scheduling virtual appointments, our research teams recommended parents to consider potential distractions throughout a given day and highlighted the importance of creating a quiet environment. We also encouraged parents to schedule appointments when a second caregiver is available to attend to other family members (such as pets and other children), leaving the participant and one parent fully attentive during the appointment. Since home environments are inevitably more distracting (Scott and Schulz, 2017), the research teams prepared parents, prior to the experiment, regarding ways to prevent potential disruptions. In all three studies, we were able to detect and handle interruptions through researcher moderation during the video conference call. However, the challenge for the researcher in these situations is conducting consistent evaluations and accommodations across subjects in order to maintain experimental control (Sheskin and Keil, 2018). We found it critical to establish a set of intervention rubrics beforehand in anticipation of various interruptions and make note of them during the appointments, as well as establishing criteria for data exclusion (e.g., if more than 10% of the trials had to be repeated or if the parent repeatedly violated protocol more than 3 times during a task). For example, in the Imitation study and the phonological awareness and vocabulary component of the AV study, individual stimuli were designed to allow representation when necessary, in order to accommodate sudden "obtrusive interferences" (e.g., significant surrounding noise in the participant's home, see **Appendix**, Table B) which were carefully defined prior to the experiment. And such accommodations were marked on the scoring sheet by the researcher.

Logistical Impacts and Cost of Online Adaptations

Reips (2002) highlighted several logistical advantages of online testing such as increased number of potential participants, lower costs, and accessibility. However, our studies did not benefit significantly in these ways. Recruitment for all three studies used pre-established participant pool databases from the University of Washington. The online procedures reduced participants' transportation costs (e.g., toll, bus fare, parking) but introduced the cost of mail delivery of equipment and/or testing materials.

Specifically, the Imitation and Reading studies involved providing electronic equipment for participants. To achieve excellent control of audio recordings across participants in the Imitation study, we mailed participants audio recorders, which enabled field recordings of speech production during virtual appointments. Similarly, inherent to the Reading study's format as a longitudinal experiment with an in-home training component, ensuring access to similar equipment (i.e., touchscreen tablets) was particularly important to the study's validity.

Both studies benefited from the high level of equipment control. However, equipment handling was a cumbersome process. It required meticulous planning such as schedule forecasting and inventory monitoring. Designated personnel prepared shipments (e.g., instructions/flow charts, equipment, small gifts, return label) sent packages at postal service locations according to the appointment schedules and even personally delivered to families when necessary. Despite the increased workload and logistical complexity caused by transporting research equipment to the families, we accepted this trade-off in order to enhance quality control of data collected in natural environments. Although we acknowledge that this is not feasible for every laboratory, sending equipment gave us the opportunity to reach a population that otherwise would not have access to these studies/ interventions.

A major logistical benefit we encountered across all three studies was increased scheduling and rescheduling flexibility for both researchers and participants. The researchers' schedule was not subjected to shared laboratory venue availability. Likewise, in addition to work-from-home conditions for many of the parents and school cancellations for children, most families reported increased daytime flexibility. Often, it was easier to squeeze a one-hour virtual appointment into their schedule compared to an in-person visit with commuting and parking difficulties. Similarly, rescheduling appointments and follow-ups with the families were easier compared to previous in-person experiences. Importantly, we could reach families who would have been unable to visit the laboratory (due to distance or availability), which increased the diversity of participants in our studies.

One disadvantage associated with online experiments, as noted by Reips (2002), is a higher attrition rate, which can be addressed by incorporating financial incentives, immediate feedback, and personalization (Frick et al., 2001). We did not notice an increased rate of withdrawal compared to previous in-person studies. We attribute this to study design considerations that were taken in order to provide logistical convenience to the families, as well as financial incentives that were similar to our in-person studies.

Presentation Mode/Setup

Moving our studies online required substantial adjustments in stimulus presentation and experimental setup. For example, the online Imitation experiment involved cartoon animations that replaced the plush puppets. The end result was visual stimuli that portrayed four cartoon characters whose mouth movements corresponded to pre-recorded audio files. During the online experiment, participants were highly engaged as cartoon characters delivered auditory stimuli. The digital animation showed to be less distracting than the puppet theater setup in the original study design. The online presentation mode eliminated distractions from tangible objects while maintaining a convincing representation of a "talking animal" for children to repeat after and interact with.

In the Imitation and AV studies, cartoon characters narrated task instructions, provided pre-programmed verbal feedback/

encouragement, and indicated experimental progress to the participants. For example, the Imitation study provided “food” rewards (e.g., bananas for the monkey character) when children completed a trial, and in the AV study, a star was displayed for every block of trials. These “rewards” served as a progress bar and motivation for the children, and digital presentation offered reliable delivery and consistent timing of the instructions, stimuli, and rewards, which helped reduce unwanted influence from the researcher during facilitation of the tasks.

When presenting auditory stimuli, output levels are important. In laboratory environments, one often uses consistent and calibrated equipment and builds experiments in a virtual environment that provides certain levels of control (e.g., Python). Since there is currently no user-friendly way to run an experiment remotely in virtual environments, the AV study reimaged the experiment by using an online experiment builder. The changes following these adaptations were substantial, but not necessarily noticeable to participants. For example, the AV study required simultaneous presentation of audio and video. We wanted to ensure that potential delays caused by the participant’s computer or browser would not affect the results. Four measures were taken to assure this. First, we pre-compiled the auditory stimuli, the noise files, and the visual part of the stimulus (photo or video). This was done using ffmpeg software (Python 3.7) on the researcher’s computer. Second, these files were then reduced in file size while keeping the quality of the sound and video.² This induced a reduction in loading time. A third precaution taken to assure simultaneous presentation was implementation of a buffer screen (200 ms blank screen) before stimulus presentation. This allowed the stimulus to fully load before it needed to be presented. And lastly, we decided to have the participants’ work go directly into the experiment builder (Lab.js), since this would avoid any delays caused by online hosting platforms (e.g., Pavlovia).

Another consideration for remote presentation of auditory stimuli is that exact loudness level on the participants’ end cannot be established. When working at a supra-threshold level, as in these studies, and/or when measuring differences in performance³ between auditory stimuli with similar qualities, exact loudness levels are not essential. A similar environment across participants was created by asking participants to set a pre-recorded speech stimulus to a comfortable level and making sure they did not change the audio settings during the experiment.

A third aspect of presenting auditory stimuli is the use of headphones. Although over-ear headphones have been accepted as the gold standard for in-laboratory auditory experiments, for all three online studies, we instructed families to use speakers for all three studies, both to control for audio output variability (compared to using headphones) across devices and to allow easy incorporation of caregiver assistance.

Control of visual presentation is often encouraged. An aspect of this, when designing the experimental setup, is the

positioning of the participant, which ideally should be consistent across participants to control for artifacts related to angle, distance, etc. Thus, preset age- and task-specific guidelines could be helpful in remote assessments. In our studies, the participants were asked to sit in a comfortable chair, or on a parent’s lap, with the computer/tablet positioned on a table in front of them. The Imitation and AV studies asked, when possible, to choose a computer over a tablet and to control the size of the display to a certain degree. With these instructions, we expected the camera angle to remain steady throughout the appointments. The Reading study also had a prescient need to ensure that the presentation of text was appropriate and consistent for each study visit. Participants were tested using a tablet (either owned or provided) for study sessions in addition to practice. In doing so, we could control for font size and scroll speed that would be adversely impacted with use of a small screen (i.e., smartphone). In the case of technical glitches that prevented use of the tablets, stimuli were projected onto the participants’ computer screen with considerations made to ensure clear and legible text and visual cues.

Video/Audio Recording

Where a researcher would be sitting adjacent or opposed to the child in the in-laboratory version of all three experiments, a similar situation was created by administering these tasks *via* a video conferencing tool. Additional to the experimenter’s role, this allowed notes to be taken, questions to be answered, and technical difficulties to be addressed. The flexibility of recording options of these video conferencing tools even facilitated some aspects of our studies.

Video Camera Setup

In our studies, Zoom video conference allowed researcher-participant communication, with the stimuli and the participant visible on screen. Similar to the in-person procedures, the Imitation experiment was video- and audio-recorded. The original setup of the study had separate cameras capture the child’s face as well as the puppet show from the child’s perspective. With the online setup, the video conferencing tool offered the convenience of being able to record both angles in the same screen share view field. In the AV experiment, disabling the researcher’s camera allowed the researcher to “hide” as an observer in the background and “appear” during necessary intervention.

Audio Setup

Considering the type of measurement (i.e., formant frequencies) in the Imitation study, obtaining quality audio recording is critical to signal analysis. However, Zoom audio recordings are subjected to input setting variability and participants’ choice of microphone. These software and hardware differences can result in incomparable speech signals or missing data. Therefore, in the absence of a highly controlled recording environment and a balanced-input microphone with exacting recording settings (as available in a laboratory booth), we sent

²<https://handbrake.fr/>

³Measuring difference of percentage correct performance between AV and audio-only stimuli presentations, rather than absolute thresholds

each family a small, child-safe⁴ LENA recorder, wearable inside a LENA vest pocket for in-home audio recording during the Zoom appointment. This setup helped minimize the distraction associated with the presence of microphones/recorders and established a controlled distance between the child's mouth and the recorder. Equipped with a power switch, a record/pause button, and a simple visual feedback mechanism, the recorder was intuitive for families to operate, lowering the risk of user error such as file deletion and data loss. Additionally, all recordings were accessible only through LENA proprietary software on a researcher's computer. This helped protect participants' data security especially since the recorders had to be returned to the researcher by mail. Despite the substantive changes introduced in our logistical procedures, sending recording equipment to the participants greatly enhanced the quality of speech data collected, bearing in mind factors that are difficult to control for in-home environments.

Moreover, we acknowledge certain benefits of auditory recordings *via* video conferencing tools of online sessions as was noted in the AV and Reading studies. Occasionally, word productions were not well perceived due to Internet lags and given that this is important for tasks like "speed reading," one could not ask the participant to repeat the stimulus. However, these "glitches" were mostly absent in audio recordings, and therefore, the test could still be scored reliably.

Other Considerations

For some studies, the format (in-person or remote) does not significantly change the implementation of audio/video recording, but recordings can be more efficient when using remote conferencing tools. In the Reading study, in-person sessions required the placement of a recording device (i.e., a handheld audio recorder) near the participant during reading activities. Not only did this introduce variability of recording quality, but perception of an explicit device tends to introduce more "performing" anxiety for child participants. On the contrary, however, we found that parents often reported that recording over the video conferencing platform helped relieve children's self-consciousness because of the use of a more integrated recording device. For the at-home training sessions, there was no recording, but it was important to log participant adherence to the practice protocol. To achieve this, we implemented online quizzes *via* Microsoft Forms. This provided a simple, secure method for participants to access the quizzes as well as for the research team to track progress.

Motivation and Sustained Attention

As described by Betts et al. (2006), sustained attention and task load have a big impact on test results for children until the age of 11–12. As children mature, their performance in accuracy and reaction times improves. We conclude that for assessments involving young children, it is important to build

in attention control (e.g., catch trials), provide multiple breaks, and decrease task load, especially online.

Task Engagement

All three studies focused on designing experiments attractive to children. The Imitation and AV studies were narrated by engaging cartoon characters that served throughout the tasks and/or used in catch trials to stimulate attention. As confirmed by Rhodes et al. (2020), animations are successful in keeping children entertained during experiments. Both children and caregivers provided feedback that these adaptations made the experiments motivating. The Reading study motivated children by choosing reading passages from a variety of topics of interest to children. But more importantly, for struggling readers, a persistent challenge is creating aids that are instructive and fun, given how taxing and frustrating reading is for this demographic. The tool *Sound it Out* was designed using evidence-based practice for reading instruction, but with an element of digital whimsy to help readers decode challenging words.

Participant Motivation

In addition to having engaging study designs, motivation can be increased by paying/rewarding subjects (Nussenbaum et al., 2020), as oftentimes, human subject payments are lower for online studies. A financial reward that is communicated to the participant before the start of the experiment or bonus rewards earned by performance can be motivators to complete longer tasks and maintain attention (Nussenbaum et al., 2020). For the studies described, participants were given an online gift card, not based on performance, with an amount similar to that provided for in-person visits. Additionally, the youngest participants received a prize toy resembling one of the cartoon characters featured in the task.

Attention Maintenance

Attention maintenance is also crucial in child studies. In all three studies, tasks were broken into sections, which allowed children to take breaks. Longer breaks were provided in between tasks. Most children were sufficiently motivated to continue without many breaks, but the opportunities were explicitly offered and even encouraged to those showing waning motivation. Particularly, the AV task had two attention mechanisms built in. First, the cartoon character would appear randomly as catch trials to measure cross-modal attention. Second, general attention was measured by including random answer options. In this 4AFC task, children picked from four answer options. All stimuli were consonant-vowel-consonant words. During the 4AFC presentation, children could pick from the goal stimulus (presented earlier in the audio-only, visual-only or audiovisual modality; e.g., sun), a minimal pair alternative (having one different consonant; e.g., run), an alternative with only the same vowel (e.g., gum), and one with no relationship to the stimulus in meaning or form (e.g., pink). We would not expect children to pick this random answer, unless they did not pay attention to the trial or fail to comprehend task instructions.

⁴LENA recorders are child safe, meeting the United States and international safety standards for electronics and toys (see www.lena.org/faqs)

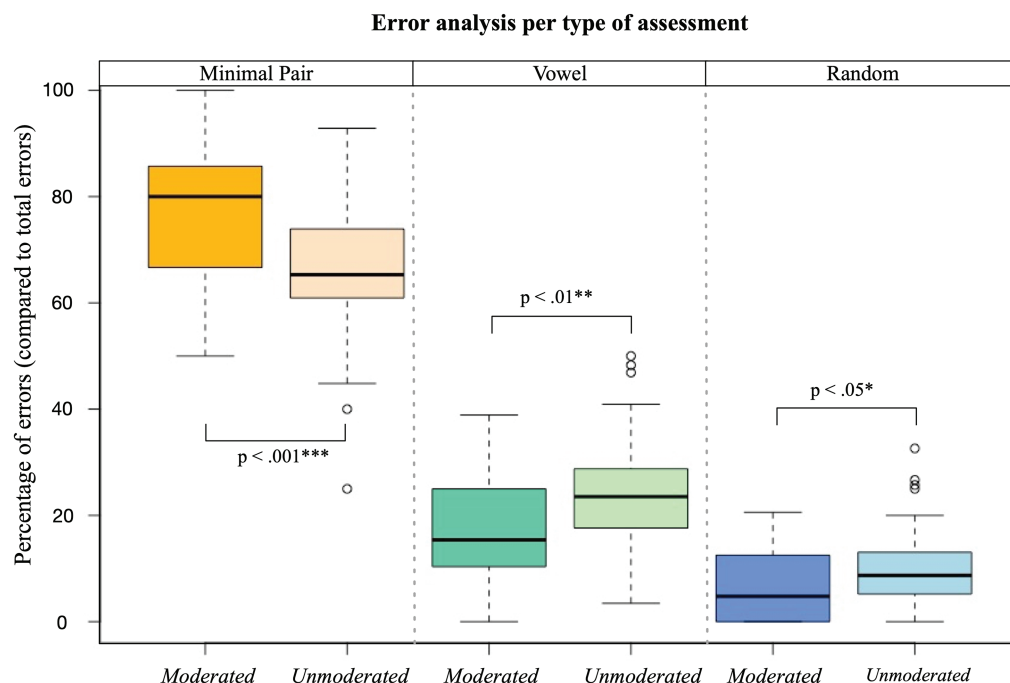


FIGURE 4 | Comparison of a moderated ($N=37$) and unmoderated ($N=47$) version of the AV task in 6-to 7-year-olds. The expected error pattern minimal pair > vowel > random errors is shown in both tasks, but more distinct for the moderated task. Random errors, and there for lack of attention is significantly higher in the unmoderated task. Thick horizontal lines represent medians, boxes represent interquartile ranges, and whiskers represent range, excluding outliers. Outliers are defined as values falling more than 1.5 x below or above the 25th and 75th percentiles, respectively, and are shown as circles. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Because all children were trained to criterion, we believe that random errors could be attributed to a lack of attention. We have facilitated this AV task moderated ($N=37$) and assessed the same task without researcher moderation in a similar group of children ($N=47$, age: 6-to 7-year-olds), as part of a bigger study. As presented in **Figure 4**, in both moderated and unmoderated assessment, children showed the expected pattern, where most errors were minimal pairs, followed by vowel words and the least responses were random words. We found that this pattern was significantly more distinct for the moderated assessment in every category. Results showed attention in the moderated task was maintained and random errors stayed low throughout the task ($M=6.63\%$, $SD=7\%$). A certain level of errors was expected, since we know attention is still developing in this population (Betts et al., 2006). In the unmoderated task, children made significantly more ($t=-2.26$, $p=0.03^*$) random errors ($M=10.14\%$, $SD=7\%$). This suggests consideration of using moderation or not when developing an online task, depending on the question asked.

Validity of Online Adaptations

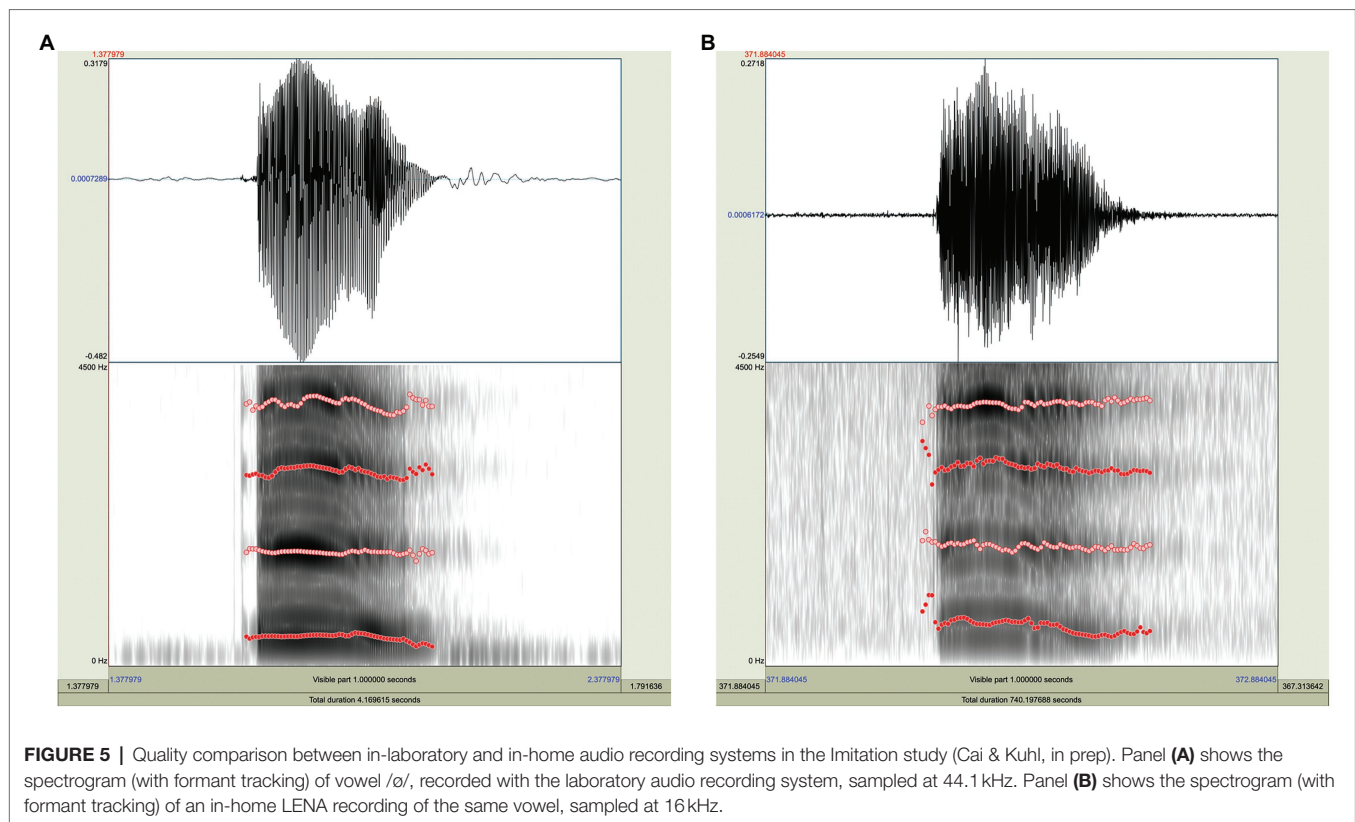
As Whitehead (2007) formulates, in situations where one uses measures online that were initially developed as paper and pencil materials, it is important to demonstrate the equivalence when one wants to interpret these similarly. Confirming existing behavioral norms online for widely used behavioral assessments would greatly benefit this process. More and more studies

designed for online testing start to confirm the possibility of getting highly reliable results online in adults (Crump et al., 2013) and children (Sheskin and Keil, 2018), even when the task is pretty different from the initial measure (Yeatman et al., 2021).

Given the nature of virtual assessments, certain factors concerning unequal audio/visual display and environmental differences were beyond our control while facilitating tasks online. However, in order to validate our remote data collection procedures, we were able to establish in-person and online comparisons within several measures critical to each study.

Validity Measures in the Imitation Study

As mentioned, the collection of speech data in the Imitation study benefited from LENA recorders' compactness, usability, and security features. However, due to the design rationale behind LENA's hardware and software systems – intending to capture day-long talk at a time, its recording quality is one 16-bit channel at a 16kHz sample rate (Ford et al., 2008), much lower than the 44.1kHz sample rate common to professional audio recordings for speech analysis. To determine whether LENA recorders were suited for this study, we tested the in-home setup and compared LENA recordings with laboratory audio samples and observed that, despite an expected lower quality in LENA recordings – associated with lower sample rates and higher background noise in natural environments – the vowel formants (i.e., the outcome measures



in the study) were equally identifiable in both sets of recordings (see Figure 5).

Validity Measures in the Audiovisual Study

Norm-referenced behavioral tasks like vocabulary tasks (e.g., EVT) are extremely valuable in developmental research, especially when researchers are specifically interested in these skills for the target group of participants. This allows the researcher to assure they have a representative group to test their specific hypothesis, and it also allows comparisons with a bigger group of children of the same age or skill level. Since there is currently little information about implementing norm-referenced tests online, a comparison from the AV study of in-person versus moderated online assessment of the EVT is shown below.

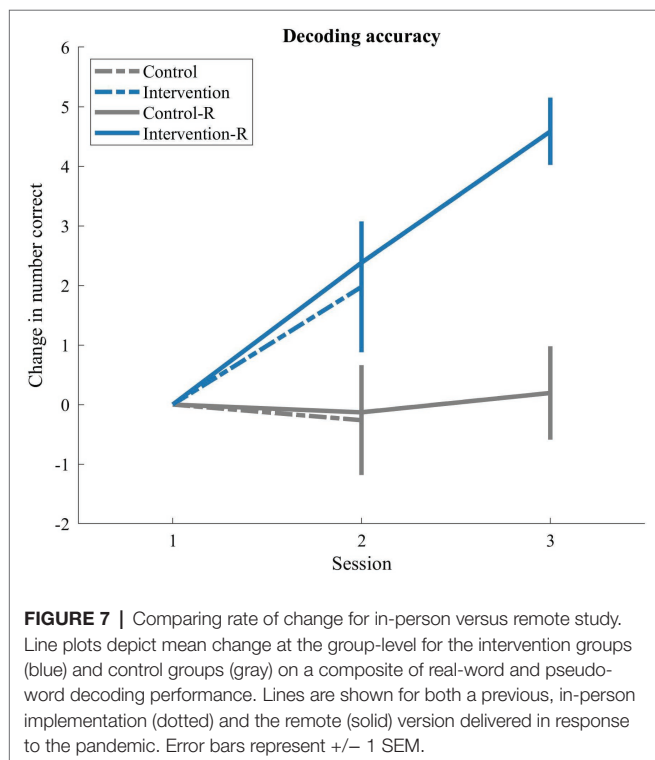
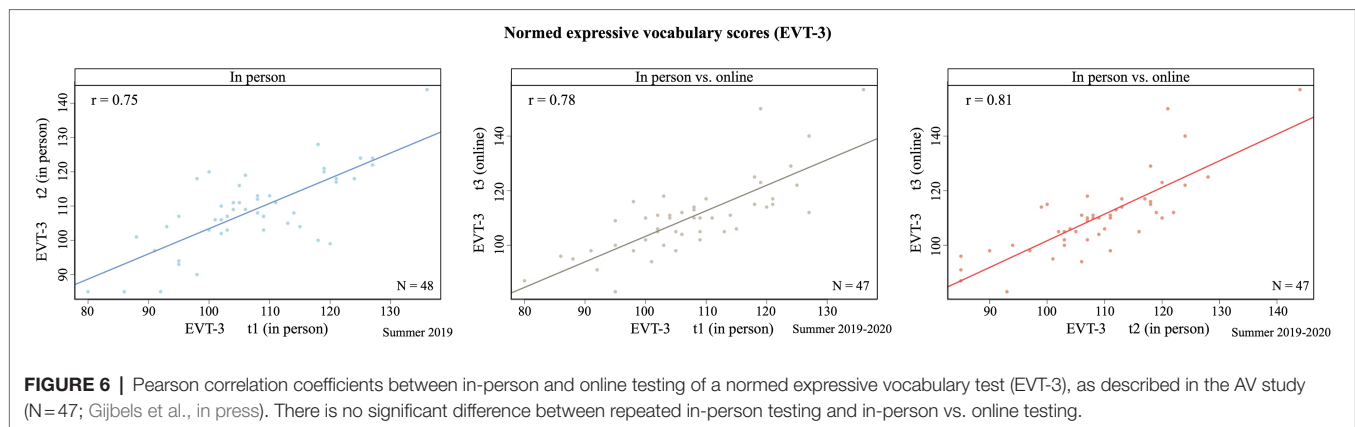
Some adaptations needed to be made to move the Expressive Vocabulary task online. Verbal instructions were given (over Zoom) following the assessment manuals, *via* a slideshow instead of the booklet. For some tasks, where children normally would have to point to a picture, the online study required them to verbalize the stimulus or the number/color attached to the picture. For children who could not do this, they were asked to point to the picture on the screen and have the caregiver verbalize it.

Since there are no data published to date confirming the use of norm-referenced scores for online assessments, we decided to interpret raw scores. This allowed comparing results between children and tasks without overcomplicating data interpretation. Nonetheless, we made a start to validate our results by doing a meta-analysis of in-person and online versions of the same

measure. Participants from the online AV study had completed the same vocabulary task (EVT) as part of an in-laboratory study in the summer of 2019. The task was assessed two times (different versions) in-person, with a 3-to 4-week separation. These children did the first version of the test again online in June 2020. The online assessment was facilitated by a trained research assistant and was conducted as similarly as possible to in-person testing. The child, caregiver, and researcher sat in front of their computers with cameras and microphones enabled, and digital scans of the materials were presented in the same way as instructed in the manual, *via* screen sharing. We found a Pearson correlation coefficient of 0.75 between the normed/standard scores of the two in-person assessments. A relationship of 0.78 was between the first (in-person) and third (online) assessment and 0.81 between the second and third assessment (see Figure 6). The correlations between the time points had no statistically significant difference ($p > 0.05$), indicating that moderated online assessment of a standardized test like this expressive vocabulary test can be reliable.

Validity Measures in the Reading Study

As previously discussed, a primary concern when the Reading study moved to a remote implementation was the ability of a virtual training program for *Sound it Out* to provide comparable benefits to those observed in an earlier, proof-of-concept study (Donnelly et al., 2020a). To our knowledge, the efficacy of remote literacy has not been explored; however, previous work in early childhood language development has shown a significant



advantage of in-person learning (Kuhl et al., 2003). Moreover, a recent survey of U.S. teachers observed that most teachers did not feel they were able to deliver the same quality of instruction when using online platforms in response to emergency school closures (Ladendorf et al., 2021).

Contrary to this concern, the Reading study demonstrated comparable-to-enhanced response in comparison with the previous, in-person iteration. As depicted in **Figure 7**, the rates of change observed after the first, two-week period of training (session 2) for the remote study (solid lines) are parallel to a similar two-week period in the previous study. Data at this shared time point indicate no significant difference between implementations for both control, $t(42) = -0.02$, $p = 0.99$, and intervention participant, $t(69) = -0.63$, $p = 0.53$, groups. Although future research is needed to determine validity, these data

suggest the significant potential for remote literacy training in the context of early childhood research.

Surprises

As much as we anticipated and prepared for obstacles associated with remote testing (e.g., instructing families to charge or connect their devices to power, conducting A/V testing at the beginning of appointments), occasional issues surfaced in the studies. For example, instead of the recommended device types, one family from the Imitation study used a Kindle tablet and needed to troubleshoot sound settings throughout the appointment due to unstable audio projection. Seldom, but present in all three studies, researchers encountered incidents where participants were disconnected mid-session either due to connection instability or low battery levels.

Overall, adopting the recorder-in-vest setup (see Section “Audio Setup”) resulted in reliable formant analysis in the Imitation study. However, because a few of the participants were not in compliance with wearing the vest, parents had to hold the recorder near the child. In these rare cases, we noticed a few instances of clipping, which is a distortion to an auditory signal when it exceeds the sensor’s constraints on the measurable range of data. In other words, the recorder could have been too close to the child’s mouth, resulting in speech input being too loud for the device.

Additionally, auditory filters and signal-to-noise adjustments on Zoom introduced additional confounds to speech tasks. For example, in the Imitation study, LENA recorders helped the researcher discover rare incidents where caregivers violated our guidelines for caregiver involvement and assisted the child during the imitation task by whispering the sounds. Such knowledge is crucial for data analysis. However, this is often undetectable over Zoom due to its background noise suppression feature. Additionally, auditory misperceptions were observed. For example, a very few participants produced /hi/ when /i/ stimuli were presented to them. Such misperception was not present in our in-person pilot work, and we suspect this to be caused by variability among audio devices and sound settings across participants. We note that the rare instances of misperception occurred only in trials containing the stimulus /i/, and vowel productions in a /h/–onset context have been shown to be virtually identical to those observed in isolation

(Kieft and Nearey, 2017). As a future step, we will explore the option of using experiment builders (as in the AV study) to deliver the stimuli for better control over the variability in audio signal transmissions.

Another data collection-related surprise occurred in the AV study. Visual stimuli included both videos and images. Because these types of stimuli were among our measures of interest, we did not draw attention to them during instruction. Occasional feedback was received about online presentations “not working” because the video seemed to have frozen. We believe this was caused by the caregivers’ realization that technical issues such as choppy videos can occur with studies online, and we suspect participants would question these occurrences less in the laboratory.

In general, we observed that children were more comfortable working from home. Although we initially thought this would lead to more distractions, participants were often less distracted by their familiar home environment than by the “new” laboratory surroundings as experienced in previous studies or pilot phases. Furthermore, it was nice to share this “from home” experience with children we had been working with before in the laboratory – for example, children loved to show their new toys or pets, which created a positive and comfortable environment for the experiments.

DISCUSSION

In this section, we will first suggest some guiding principles derived from our implementations of the three online studies in order to aid developmental scientists seeking to carry out future online studies. Next, we will look deeper into the current limitations of online behavioral testing involving children as well as some resources and future improvements needed to move the field forward online.

Guiding Principles Generated From the Three Studies

The studies discussed in this paper differed in research questions explored and age groups involved. However, commonalities and differences among the studies lend themselves to suggesting the following guiding principles for future online developmental studies.

In general, remote consent procedures can take place over secure online portals. But the downside of solely obtaining (electronic) signatures online is the lack of explicit opportunity for participants/caregivers to raise questions and/or concerns. We recognize that it is important to consider consent acquisition as a process rather than a product (Whitehead, 2007), especially when children are involved. Therefore, we posit that a valuable step to take is to ensure participants’ and caregivers’ understanding of informed consent through researcher moderation. This can serve to supplement written consent procedures or can occur as a separately documented process to replace text-based consent forms.

The degree of caregiver involvement is typically determined by the age group and the complexity of equipment manipulation.

Involving caregivers of younger participants in our studies required intentional efforts to ensure that they followed the research protocol closely to avoid introducing unwanted interference. Clear communication of research protocols prior to the appointment is crucial in establishing desired caregiver involvement. Additionally, we experienced that it was helpful to provide families visualizations of experimental procedures or scripts of approved caregiver encouragements. Therefore, in addition to a carefully designed protocol, we believe that these steps could help minimize the confounding risk of caregiver interference. Although the level of caregiver involvement differed by age, technical support was critical for all three studies. When active manipulation of technical devices (e.g., mouse clicks) is required by the children, it can be helpful to objectively assess technical proficiency of the child during a training session, and based on the outcome, decisions can be made regarding caregivers’ assistance in technical manipulations.

During data acquisition, it is crucial to generate and deliver consistent stimuli across subjects. However, in remote studies containing visual and auditory stimuli, it is more complicated to ensure this. Each of the three studies attempted to control for the quality of stimuli delivery in their own way, from screen sharing pre-recorded sets of cartoon animations, to providing participants with designated software. Generating and delivering testing materials using experiment builders would be a favorable option as the automation of stimulus delivery has been reported to reduce the workload of the researcher during the task, lowering the chance of human error (Rhodes et al., 2020).

Related to this, we encourage future studies to carefully evaluate the benefits and costs of providing research equipment to the participants following targeted research questions and data types. In our studies, we made logistical decisions based on task designs and resources available. In the Imitation study, mailing LENA recorders and vests to all participating families was a sensible and effective choice because consistency of speech recordings across participants was critical to the experiment. And it was to our unique logistical advantage that we could use existing resources (i.e., the LENA recorders) which happened to be participant-friendly, since the families had participated in our previous research using the same device. Since the Reading study required the use of a tablet-based app, there was a need to mail a tablet to participants who had no access to one. The goal of the study was to provide an intervention/aid for a population that needs help with literacy development. When only including families that own a tablet, a large portion of this population would have been excluded. For the AV study, it was not necessary to send equipment due to the type of data measured. This study took a different approach in experimental control where, through the use of an experimenter builder, general cross-subject consistency in participants’ visual and auditory perception was achieved.

Another helpful measure to ensure experimental control for online developmental studies is researcher moderation. Although most online behavioral procedures can be automated,

it is beneficial to control for unexpected changes in the environment, allowing for impromptu adjustments and extra technical support. We suggest from the findings in the AV studies that moderation could help improve participants' attention. The researcher can be aware of any decline in participants' attention and suggest a break or introduce adequate motivators. Additionally, researcher moderation allowed participants and their caregivers to ask questions during the consent procedure and ensured that no data would be lost due to invalid consent/assent procedures. Finally, we believe that the personal connection we established with the participants through moderation was beneficial to lowering the attrition rate and helped sustain participants' attention.

Last but not least, due to the variability and complexity of study designs in developmental research, validation of online methods in this field often stays specific to each study. We believe a potential solution may be to carry out a study design both in-person and remotely during the initial pilot phase and assess the validity of the online study design by comparing pilot results. Moreover, when designing an online study or converting an in-person study to virtual environments, it is consequential to identify areas of adaptation and define the purpose of each adaptation. Meticulous deliberation and systematic documentation of such decisions would maximize the comparability between data collected in-person and remotely and could benefit future replications of the study within or between laboratories.

Toward a Future of Remote, Moderated Studies of Early Childhood Development Generalizability/Reliability

Researchers desire highly controlled study designs and environments for accurate experimental measures, sometimes at the cost of results generalizability. Virtual settings promote a natural environmental variability, which could increase ecological validity and generalizability (Laugwitz, 2001; Reips, 2002). Depending on the type of research, exploring previously documented findings in naturalistic settings can be useful. Of course, this varies by types of research. As Reips (2002) suggests, behavioral research that is conducted on topics with no relation to computer-mediated communication might make interpretation more selective instead of more generalizable.

With regard to reproducibility of research findings, noise in measurement and contextual factors may compromise reproducibility (Frank et al., 2017). Online methods could make it easier to share digital stimuli, and participants' environmental control would be comparable from study to study. As online research tasks need to be more automated, participants do not heavily depend on researchers' involvement in stimulus delivery, reducing interactive bias (Rhodes et al., 2020).

Although Krantz and Dalal (2000) claim equal external validity between in-person and remote testing, currently the comparison between the validity of data collected in-person vs. online is incomplete and needs further evidence. Kim et al. (2019) concluded that, depending on the measure of interest, data collected in-person and online can be comparable or

equivalent. They found that replicating in-person studies online did not have a noticeable impact on participants' response accuracy but affected their reaction time. Reips (2002) added that individual hardware differences, Internet connection, and background running programs can have an effect on data collection consistency across participants and that validity and reliability of online experiments will need to be expanded in the future.

Inclusive, Equitable Research

It has been reported that most in-laboratory developmental studies recruit children from areas surrounding universities (Henrich et al., 2010). While online recruitment opens doors to broaden participant recruitment and diversify the subject pool, the diversity is not guaranteed and the change will not happen overnight. Future work is needed to identify barriers to reach diverse populations. According to the National Center for Education Statistics, in 2016, over 80% of the households in the United States have access to the Internet, and in 2018, 90% of the U.S. population owned a desktop computer, laptop, or tablet. This number is increasing every year. Although the numbers with access to technology are high and increasing, there still exist barriers and inequities for online research in a large group of the population, which is associated with lack of access to these resources among certain populations (Neuman and Celano, 2006; Jenkins, 2009). As these are often families of lower income, lower education levels or minorities, online research may bias toward recruiting specific groups of the population, similar to in-person research. Furthermore, research might not be inviting to these hard-to-reach populations. Shaghaghi et al. (2011) point out that it is only possible to reach a wider population if you make active social, cultural, or behavioral adjustments to make the research more meaningful and accessible.

Resources

Converting studies online can seem intimidating for many because of the adjustments that need to be made. However, the changes can be quite positive. At times, crises can force adaptation and encourage advancements. Even beyond the pandemic, we believe that online developmental research can be as valuable or even more valuable than in-person research when thoughtful adjustments and considerations are made.

Although we initially felt there was little support for online adaptations from the developmental science literature, we discovered platforms such as Lab.js, Gorilla, and Pavlovia, as well as task forces such as "The Acoustical Society of America's Task Force on Remote Testing," which are investing immensely in support systems for researchers interested in virtual studies. Furthermore, other researchers running into similar difficulties while developing online behavioral experiments are starting to report their experiences (e.g.; Sauter et al., 2020). We are hopeful that this trend will continue, and as a result, future studies moving online will benefit from access to more developed systems to start collecting online data with confidence.

CONCLUSION

Similar to diverse laboratory-based experimental designs, online methodologies are specific to individual research questions. The three studies mentioned in this paper employed different methods and encountered problems unique to their study design. We hope our experiences will be informative for future remote studies beyond the impact of the COVID-19 pandemic.

We believe by adjusting our developmental research methods from traditional in-person settings to an online format and by acknowledging all the changes needed to be made, our developmental work is as valuable as it would have been in-person. All children could participate from a familiar environment at a time that worked for both them and the researcher, without having to make concessions and, for example, arrive at the laboratory after a long day of school, activities, and driving. Testing from home can positively impact general attention and comfort for children. In our observations, many of our participants wanted to share their world (e.g., toys, pets) with the researcher and were highly motivated to participate. Data collection procedures felt more natural and comfortable for them because they completed the tasks in their home environment. Additionally, we recognize that part of the reason for the ease of our recruitment and the high compliance from our participants could be that we had established strong rapport with most of the participants and their caregivers from previous studies.

All experimental control that would be routine in a laboratory environment had to be reevaluated and adjusted for online testing, which led to carefully considered and documented protocols. This, in combination with the automation of the research tasks, may make it easier for others to replicate our analyses and findings. As our observations (*via* moderation) and results show consistency over participants and home environments, we believe we succeeded in tackling what we initially observed as the most challenging parts of remote developmental work. This goes from finding platforms and technical support to move the experiment online, to control of the participant's environment and even logistical issues. However, by no means does this paper attempt to license one "correct" set of rules all online developmental research studies should follow. Instead, by sharing our experiences, we would like to call attention to the need for reported evidence of adapted remote studies in this field. We believe that more experiences of online developmental studies remain to be had and shared.

REFERENCES

- Abrahamsson, N., and Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Stud. Second. Lang. Acquis.* 30, 481–509. doi: 10.1017/S027226310808073X
- Benton, L., Vasalou, A., Berkling, K., Barendregt, W., and Mavrikis, M. (2018). A critical examination of feedback in early reading games. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.
- Betts, J., McKay, J., Maruff, P., and Anderson, V. (2006). The development of sustained attention in children: the effect of age and task load. *Child. neuropsychol. j. normal. abnormal develop. childhood. adolescence*. 12, 205–221. doi: 10.1080/09297040500488522

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Washington Human Subjects Division. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

LG, RC, PMD, and PKK contributed to conception and design of the studies. LG, RC, and PMD executed and analyzed the studies. LG and RC wrote the first draft of the manuscript. LG, RC, and PMD wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was funded by NSF BCS 1551330, NIH NICHD R01HD09586101, NICHD R21HD092771, Microsoft Research Grants and a Jacobs Foundation Research Fellowship to Jason D. Yeatman and by the Overdeck Family Foundation, the University of Washington Institute for Learning & Brain Sciences Ready Mind Project. The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.740290/full#supplementary-material>

- Blok, H., and Oostdam, R., Otter, M. E., and Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: a review. *Rev. Educ. Res.* 72, 101–130. doi: 10.3102/00346543072001101.
- Bohner, G., Danner, U. N., Siebler, F., and Samson, G. B. (2002). Rape myth acceptance and judgments of vulnerability to sexual assault: an Internet experiment. *Experimental psychology*. 49, 257–269. doi: 10.1026/1618-3169.49.4.257
- Cheung, A. C. K., and Slavin, R. E. (2011). The effectiveness of education technology for enhancing reading achievement: a meta-analysis. *Best. Evidence. Encyclopaedia*. 97, 1–48.
- Christ, T., Poonam, A., and Yu, L. (2018). Technology integration in literacy lessons: challenges and successes. *Literacy. Res. Instruc.* 58, 1–18. doi: 10.1080/19388071.2018.1554732

- Christiner, M., and Reiterer, S. M. (2013). Song and speech: examining the link between singing talent and speech imitation ability. *Front. Psychol.* 4:874. doi: 10.3389/fpsyg.2013.00874
- Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS one.* 8:186. doi: 10.1371/journal.pone.0057410
- Donnelly, P. M., Gijbels, L., Larson, K., Matskewich, T., Linnerud, P., Kuhl, P. K., et al. (2020b). A symbolic annotation of vowel sounds for emerging readers. *PsyArXiv*. doi: 10.31234/osf.io/akjdr
- Donnelly, P. M. K., Larson, T. M., and Yeatman, J. D. (2020a). Annotating digital text with phonemic cues to support decoding in struggling readers. *PLoS One* 15:e0243435. doi: 10.1371/journal.pone.0243435
- Duffy, M. E. (2002). Methodological issues in web-based research. *J. nursing. Scholarship.* 34, 83–88. doi: 10.1111/j.1547-5069.2002.00083.x
- Ford, M., Baer, C., Xu, D., Yapnel, U., and Gray, S. (2008). *The LENA language environment analysis system: audio specifications of the DLP-012 (Technical Report LTR-03-2)*. Boulder, CO: LENA Foundation.
- Fort, M., Spinelli, E., Savariaux, C., and Kandel, S. (2012). Audiovisual vowel monitoring and the word superiority effect in children. *Int. J. Behav. Dev.* 36, 457–467. doi: 10.1177/0165025412447752
- Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., et al. (2017). A collaborative approach to infant research: promoting reproducibility, best practices, and theory building. *Infancy. official j. Int. Soc. Infant. Stud.* 22, 421–435. doi: 10.1111/inf.12182
- Franken, M. K., Hagoort, P., and Acheson, D. J. (2015). Modulations of the auditory M100 in an imitation task. *Brain. Language.* 142, 18–23. doi: 10.1016/j.bandl.2015.01.001
- Frick, A., Bächtiger, M., and Reips, U. (2001). "Financial incentives, personal information and drop out in online studies," in *Dimensions of Internet Science*. eds. U.-D. Reips and M. Bosnjak (Lengerich: Pabst), 209–219.
- Ghazi-Saidi, L., and Ansaldo, A. I. (2017). Second language word learning through repetition and imitation: functional networks as a function of learning phase and language distance. *Front. Hum. Neurosci.* 11:463. doi: 10.3389/fnhum.2017.00463
- Gibson, E., and Twycross, A. (2008). Getting it right for children and young people's health care services. *J. Clin. Nurs.* 17, 3081–3082. doi: 10.1111/j.1365-2702.2008.02644.x
- Gijbels, L., Yeatman, J. D., Lalonde, K., and Lee, A. K. (in press). Audiovisual speech processing in relationship to phonological and vocabulary skills in first graders. *J. Speech Lang. Hear. Res.* doi: 10.1177/0265659018793697
- Grant, K. W., and Bernstein, J. G. W. (2019). "Toward a model of auditory-visual speech intelligibility," in *Multisensory Processes* (Cham: Springer International Publishing), 33–57.
- Grant, A., Wood, E., Gottardo, A., Evans, M. A., Phillips, L., and Savage, R. (2012). Assessing the content and quality of commercially available reading software programs: do they have the fundamental structures to promote the development of early reading skills in children? *NHSA Dialog* 15, 319–342. doi: 10.1080/15240754.2012.725487
- Guernsey, L., and Levine, M. H. (2015). *Tap, Click, Read: Growing Readers in a World of Screens*. San Francisco: Jossey-Bass.
- Gweon, H., Sheskin, M., Chuey, A., and Merrick, M. (2020). Video-chat studies for online developmental research: Options and best practices. *Social Learning Lab Webinar*. Available at: http://sll.stanford.edu/docs/Webinar_materials_v2.pdf (Accessed May 30, 2021).
- Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., and Hilbig, B. E. (2020). Lab.js: a free, open, online study builder. *Behav. Res. Methods*, 1–18. doi: 10.5281/zenodo.597045
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world?. *The Behavioral and brain sciences.* 33, 61–135. doi: 10.1017/S0140525X0999152X
- Hewson, C. M., Laurent, D., and Vogel, C. M. (1996). Proper methodologies for psychological and sociological studies conducted via the internet. *Behav. Res. Methods. Instrum. Comput.* 28, 186–191. doi: 10.3758/BF03204763
- Holt, R. F., Kirk, K. I., and Hay-McCutcheon, M. (2011). Assessing multimodal spoken word-in-sentence recognition in children with normal hearing and children with Cochlear implants. *J. Speech Lang. Hear. Res.* 54, 632–657. doi: 10.1044/1092-4388(2010/09-0148)
- Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., and Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) learners: Behavioral predictors and neural substrates. *Brain Lang.* 127, 366–376. doi: 10.1016/j.bandl.2012.11.006
- Jenkins, H. (2009). *Confronting the Challenges of Participatory Culture: Media Education for the 21st Century*. Cambridge, MA: The MIT Press.
- Jerger, S., Damian, M. F., Spence, M. J., Tye-Murray, N., and Abdi, H. (2009). Developmental shifts in children's sensitivity to visual speech: a new multimodal picture-word task. *J. Exp. Child Psychol.* 102, 40–59. doi: 10.1016/j.jecp.2008.08.002
- Jerger, S., Damian, M. F., Tye-Murray, N., and Abdi, A. K. (2014). Children use visual speech to compensate for non-intact auditory speech. *J Exp Child Psychol.* 126, 295–312. doi: 10.1016/j.jecp.2014.05.003
- Kieffe, M., and Nearey, T. M. (2017). Modeling consonant-context effects in a large database of spontaneous speech recordings. *J. Acoust. Soc. Am.* 142:434. doi: 10.1121/1.4991022
- Kim, J., Gabriel, U., and Gyga, P. (2019). Testing the effectiveness of the internet-based instrument PsyToolkit: a comparison between web-based (PsyToolkit) and lab-based (E-prime 3.0) measurements of response choice and response time in a complex psycholinguistic task. *PLoS One* 14:e0221802. doi: 10.1371/journal.pone.0221802
- Krantz, J. H., and Dalal, R. (2000). "Validity of web-based psychological research," in *Psychological Experiments on the Internet*. ed. M. H. Birnbaum (San Diego, CA: Academic Press), 35–60.
- Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., and Couper, M. (2004). Psychological research online. *Am. Psychol.* 59, 105–117. doi: 10.1037/0003-066X.59.2.105
- Kuhl, P. K., and Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behav. Dev.* 7, 361–381. doi: 10.1016/S0163-6383(84)80050-8
- Kuhl, P. K., Tsao, F. M., and Liu, H. M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proc. Natl. Acad. Sci.* 100, 9096–9101. doi: 10.1073/pnas.1532872100
- Ladendorf, K., Muehsler, H., Xie, Y., and Hinderliter, H. (2021). Teacher perspectives of self-efficacy and remote learning due to the emergency school closings of 2020. *Educ. Media Int.*, 1–21. doi: 10.1080/09523987.2021.1930481
- Lalonde, K., and McCreery, R. W. (2020). Audiovisual enhancement of speech perception in noise by school-age children who are hard of hearing. *Ear Hear.* 41, 705–719. doi: 10.1097/AUD.0000000000000830
- Lalonde, K., and Werner, L. A. (2021). Development of the mechanisms underlying audiovisual speech perception benefit. *Brain Sci.* 11:49. doi: 10.3390/brainsci11010049
- Lambert, V., and Glacken, M. (2011). Engaging with children in research. *Nurs. Ethics* 18, 781–801. doi: 10.1177/0969733011401122
- Laugwitz, B. (2001). "A web experiment on color harmony principles applied to computer user interface design," in *Dimensions of Internet Science*. eds. U.-D. Reips and M. Bosnjak (Lengerich, Germany: Pabst Science), 131–145.
- McTigue, E. M., Solheim, O. J., Zimmer, W. K., and Uppstad, P. H. (2020). Critically reviewing GraphoGame across the world: recommendations and cautions for research and implementation of computer-assisted instruction for word-reading acquisition. *Read. Res. Q.* 55, 45–73. doi: 10.1002/rrq.256
- Neuman, S. B., and Celano, D. (2006). The knowledge gap: implications of leveling the playing field for low-income and middle-income children. *Read. Res. Q.* 41, 176–201. doi: 10.1598/RRQ.41.2.2
- Nussenbaum, K., Scheuplein, M., Phaneuf, C., Evans, M., and Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra. Psychology.* 6. doi: 10.1525/collabra.17213
- Reips, U. (2001). The web experimental psychology lab: five years of data collection on the internet. *Behav. Res. Methods. Ins. Comp.* 33, 201–211. doi: 10.3758/BF03195366
- Reips, U. (2002). Standards for internet-based experimenting. *Exp. Psychol.* 49, 243–256. doi: 10.1026/1618-3169.49.4.243
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Richardson, U., and Lyytinen, H. (2014). The GraphoGame method: the theoretical and methodological background of the technology-enhanced learning

- environment for learning to read. *Hum. Technol.* 10, 39–60. doi: 10.17011/ht/urn.201405281859
- Ronimus, M., and Lyytinen, H. (2015). Is school a better environment than home for digital game-based learning? The case of GraphoGame. *Hum. Technol.* 11, 123–147. doi: 10.17011/ht/urn.201511113637
- Ross, L. A., Molholm, S., Blanco, D., Gomez-Ramirez, M., Saint-Amour, D., and Foxe, J. J. (2011). The development of multisensory speech perception continues into the late childhood years. *Eur. J. Neurosci.* 33, 2329–2337. doi: 10.1111/j.1460-9568.2011.07685.x
- Sauter, M., Draschkow, D., and Mack, W. (2020). Building, hosting and recruiting: a brief introduction to running Behavioral experiments online. *Brain Sci.* 10:251. doi: 10.3390/brainsci10040251
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (part 2): assessing the viability of online developmental research, results from three case studies. *Open mind* 1, 15–29. doi: 10.1162/OPMI_a_00002
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Shaghghi, A., Bhopal, R. S., and Sheikh, A. (2011). Approaches to recruiting 'hard-to-reach' populations into research: a review of the literature. *Health Promot. Perspect.* 1, 86–94. doi: 10.5681/hpp.2011.009
- Sheskin, M., and Keil, F. (2018). A video chat platform for developmental research. *TheChildLab.com*. Available at: <https://doi.org/10.31234/osf.io/rn7w5> (Accessed June 22, 2021).
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Soe, K., Stan, K., and Juvenna, M.C. (2000). Effect of Computer-Assisted Instruction (CAI) on Reading Achievement: a Meta-Analysis. *Pacific Resources for Education and Learning*. Available at: <https://eric-ed-gov.offcampus.lib.washington.edu/?id=ED443079> (Accessed June 19, 2021).
- Stetter, M. E., and Hughes, M. T. (2010). Computer-assisted instruction to enhance the reading comprehension of struggling readers: a review of the literature. *J. Spec. Educ. Technol.* 25, 1–16. doi: 10.1177/016264341002500401
- Whitehead, L. C. (2007). Methodological and ethical issues in internet-mediated research in the field of health: an integrated review of the literature. *Soc. Sci. Med.* 65, 782–791. doi: 10.1016/j.socscimed.2007.03.005
- Williams, K. (2014). *Phonological and Print Awareness Scale*. Torrance, CA: WPS Publishing.
- Williams, K. T. (2019). Expressive Vocabulary Test [Measurement instrument]. 3rd Edn. Bloomington, MN: NCS Pearson.
- Wolf, M., Gottwald, S., Galyean, T., Morris, R., and Breazeal, C. (2014). "The reading brain, global literacy and the eradication of poverty," in *Bread and Brain, Education and Poverty*. eds. A. Battro, I. Potrykus, and M. S. Sorondo (Vatican City: Pontifical Academy of Sciences), 1–22.
- Yeatman, J. D., Tang, K. A., Donnelly, P. M., Yablonski, M., Ramamurthy, M., Karipidis, I. I., et al. (2021). Rapid online assessment of reading ability. *Sci. Rep.* 11:6396. doi: 10.1038/s41598-021-85907-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Gijbels, Cai, Donnelly and Kuhl. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

APPENDIX

Table A Summary description of the three discussed studies. The table provides more detail about the area of expertise, the equipment used, and explains per study degree of moderation, informed consent, caregiver involvement, logistical impact, presentation mode, video and audio recording, motivation and sustained attention, data interpretation, and surprises.

Table B Example protocol for handling various types of obtrusive interferences during online facilitation of the Imitation Study. The table categorizes potential interferences that may disrupt data collection and the measures taken (both before and during virtual appointments) to address the disruptions.



Parent-Infant Interaction Tasks Adapted for Remote Testing: Strengths, Challenges, and Recommendations

Shira C. Segal* and Margaret C. Moulson

Department of Psychology, Ryerson University, Toronto, ON, Canada

OPEN ACCESS

Edited by:

Natasha Kirkham,
Birkbeck, University of London,
United Kingdom

Reviewed by:

Sabrina L. Thurman,
Elon University, United States
Fanli Jia,
Seton Hall University, United States

*Correspondence:

Shira C. Segal
shira.segal@ryerson.ca

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 June 2021

Accepted: 21 September 2021

Published: 13 October 2021

Citation:

Segal SC and Moulson MC (2021)
Parent-Infant Interaction Tasks
Adapted for Remote Testing:
Strengths, Challenges, and
Recommendations.
Front. Psychol. 12:733275.
doi: 10.3389/fpsyg.2021.733275

The closure of in-person laboratories and decreased safety of face-to-face interactions resulting from the COVID-19 pandemic jeopardized the ability of many developmental researchers to continue data collection during this time. Disruptions in data collection are particularly damaging to longitudinal studies, in which the testing of different age groups occurs on a continuous basis, and data loss at one time point can have cascading effects across subsequent time points and threaten the viability of the study. In an effort to continue collecting data for a longitudinal study on emotion development started in-person pre-pandemic, we adapted two parent-infant interaction tasks (free-play task and toy removal task) for a remote testing framework. Our procedure for pivoting these tasks to a supervised, remote online testing framework is outlined and the associated strengths and challenges of testing in this format (e.g., feasibility and implementation, testing environment and task setup validity, and accessibility, recruitment, and diversity) are critically evaluated. Considerations for applying this framework to other behavioral tasks are discussed and recommendations are provided.

Keywords: remote research methods, online testing, COVID-19, videoconferencing, infancy, parent-child interaction, zoom

INTRODUCTION

With the onset of the COVID-19 pandemic and university closures around the globe, developmental researchers largely found themselves forced to move their work activities to a remote platform. Among the many different tasks of a developmental researcher, one activity posed especially difficult challenges in the pivot to the new remote setup: data collection. Although some researchers had previously developed protocols for online testing with developmental samples (e.g., the Lookit platform; Scott and Schulz, 2017; TheChildLab.com; Sheskin and Keil, 2018), the pandemic sparked a widespread need to embrace remote testing as one of the only viable options to continue collecting data (for an example of a collaborative initiative for online testing founded during the pandemic, see ChildrenHelpingScience.com).

One particular context in which disruptions in data collection can have cascading and enduring effects is the collection of longitudinal data. In a longitudinal study, research questions are designed on the premise of having follow-up data at each time point; successful data collection is contingent upon being able to continuously test participants as they “age in” to each brief window of eligibility for participation. When working with infants, these windows of eligibility may be as short as 1–2 weeks, depending on the age requirements for participation. Losing data during the follow-up time points of a longitudinal study can render previous years’ worth of data collection and countless time offered up by families unusable. In these cases, there is an obligation to numerous participating bodies to find a way to continue a given research project. There is an obligation to the families who offered their time in the hopes of contributing to the scientific research outlined to them when they consented to enroll in a study; there is an obligation to funding agencies, who provided funds and entrusted the researcher to carry out the proposed work to completion; and on a more personal level, researchers may feel an obligation to the many different laboratory personnel who dedicated their time to helping a study run smoothly over the years, some of whom may be relying on completion of longitudinal data collection for training milestones (e.g., dissertation). The continuity of longitudinal data collection was the primary motivation behind our laboratory’s development of a novel protocol for online testing.

Our longitudinal study was a multi-method study on emotion development across the first 2 years of life. Infants were tested at 3.5, 7, 12, and 18 months. At 3.5 and 7 months, tasks included the still face paradigm (Tronick et al., 1978), ERP, a free-play interaction, and eye tracking, and at 12 and 18 months, tasks included eye tracking and a parent-infant toy removal task (Stifter and Braungart, 1995; for more detailed descriptions of the larger study see Segal and Moulson, 2020a, Segal and Moulson, 2020b, and Segal et al., 2021). Data collection began in May 2017, and when testing was shut down in March 2020, we had collected data for 78% of our target sample at 3.5 months, 50% at 7 months, 29% at 12 months, and 21% at 18 months. Many aspects of our longitudinal study were not amendable to the switch to online testing; however, we decided to resume testing in October 2020 by adapting two of our parent-infant interaction tasks for remote testing.

This report will outline how we adapted these two tasks, a toy removal task and a free-play task, for online testing. In contrast to preexisting frameworks for remote, unmoderated testing (e.g., Scott and Schulz, 2017; Rhodes et al., 2020), the current framework outlines a method of supervised remote testing in which the researcher is available to guide families through the testing procedures in real-time. This report does not present an empirical comparison across methods; rather, we present areas of considerations for researchers who may be in the preparation or planning stages of moving an in-person task to online testing. There are few guidelines available detailing this process, so our goal is to highlight methodological considerations that may be applicable to the adaptation of

other behavioral tasks, beyond the two tasks presented in this paper.

MATERIALS AND METHODS

Task Descriptions

Toy Removal Task

The toy removal task (Stifter and Braungart, 1995) was designed for measuring emotion regulation in infants, as it simulates a routine frustration-eliciting situation and provides an opportunity for researchers to measure regulatory behaviors. Our instantiation of this task consists of four phases: (1) play (1.5 min): parents and infants are provided with a toy and they are instructed to play together; (2) toy removal (2 min): parents are instructed to take the toy away and place it somewhere out of reach but still within sight of the infant. Parents are requested to refrain from speaking to or touching the infant during this time and may be provided with materials to help keep their attention directed away from their infant (e.g., questionnaire or magazine); (3) parent attention return (1 min): without returning the toy, parents are permitted to resume interacting with the infant as normal (e.g., talking and touching); and (4) toy return (1 min): parents are prompted to return the toy to the infant and to resume playing together.

Free-Play Task

This task consists of a 10-min free-play interaction between parents and infants. Parents are instructed to play with their infant as they normally do at home, and they are permitted to use any toys available to them. The interaction is video recorded.

Online Testing Procedure

Both tasks were run synchronously during an online testing session guided by a researcher over Zoom (Zoom Video Communications Inc., San Jose, CA, United States). Families who had previously participated in our longitudinal study and whose infants were approaching eligibility for the 12- or 18-month time points were contacted and provided information about the online continuation of the study. Parents were sent a consent form to review, and interested parents were provided with the option to either send back a signed consent form prior to the visit or to provide verbal consent during the testing session. To ensure consistent use of the same toy across participants during the toy removal task, a busy box toy (VTECH Busy Learners Activity Cube) was sent to families in advance using Amazon Prime shipping service. Parents were instructed not to open the toy until the testing session to ensure that it remained equally novel to each infant at the start of the task. Parents who indicated that they had a printer at home were asked to print the Depression Anxiety Stress Scale-21 (DASS-21; Lovibond and Lovibond, 1995) ahead of time (to serve as a distraction for parents during the toy removal phase), and parents who were unable to print it were reassured that the researcher would complete it with them

during the testing session. Parents were instructed not to complete the questionnaire ahead of time, as this questionnaire is sensitive in nature, and we preferred that it be completed in the presence of a researcher with clinical training to allow for debriefing.

During the testing session, the researcher first collaborated with the family in finding an optimal setup for the task within their homes, which consisted of the infant and parent seated beside each other at a table with the infant seated in a high chair. An important consideration included assisting the family in finding a location with sufficient lighting to see the infant's face during the recording (e.g., avoiding backlighting). The researcher reviewed the consent form and task instructions with families. When reviewing the instructions, parents were told that a tone would be played to indicate when to move into each phase of the toy removal task, and the researcher previewed the tone for parents. Parents who were unable to print the DASS-21 prior to the visit were instructed to keep a magazine, book, or their phone nearby to use during the toy removal phase in place of the questionnaire. For parents who opted to provide verbal consent, a standardized consent agreement was pasted into the Zoom chat, and the parent was asked to read it aloud after the recording began. The researcher started recording the task within Zoom and turned off their video and microphone for the duration of the task. During the task, the researcher observed the interaction and timed each task phase, ensuring that the tone was audible for each phase transition (i.e., briefly unmuting to play the tone). If parents required assistance during the task, the researcher was available to guide the parents through task-related or technology-related issues. The task was ended early in cases where infants exhibited consistent crying for greater than 20 s, which was the same criterion applied during in-person testing. After the task was completed, the researcher stopped the Zoom recording.

The free-play task was completed directly following completion of the toy removal task. The dyad was given a short break while the researcher explained the rationale, instructions, and required setup for the free-play task. Parents were instructed to bring the camera to a location in their home containing the toys with which the infant typically plays and to set it up so that the entire scene was viewable (i.e., full body of both participants). The researcher asked the parent to orient the infant to be facing the camera when possible (to allow for later facial coding) and provided them with the instruction to play as they normally do at home for 10 min. The researcher started a second recording and turned off their camera and microphone once again. After 10 min, the researcher stopped the recording. In the case of participants who did not have access to a printer and were therefore unable to complete the DASS-21 questionnaire during the first task, the researcher shared the questionnaire on their screen and completed it with the parent virtually. After completion of the free-play task (and DASS-21 questionnaire when necessary), families were debriefed and given a chance to ask questions about the study.

The toy that participants received in the mail for the toy removal task (VR VTECH Busy Learners Activity Cube) also served as participants' compensation for participating in the study, as families were given the toy to keep after participation, and it was of similar value to previous in-person monetary compensation.

DISCUSSION

In this report, we outlined the methodology of two behavioral interaction-based tasks adapted for online testing with infants between 12 to 24 months. Guidelines detailing how to adapt in-person tasks for online testing are scarce, and as researchers increasingly embrace remote testing, the development of frameworks designed to help researchers make this transition is well-timed. Our goal is to provide an overview of the adaptations required to modify these tasks for online testing to critically evaluate the strengths and challenges of collecting data in this format. Our considerations may not be applicable to all parent-child interaction tasks, but our hope is that our general approach of adapting the two tasks described above may serve as a starting framework for other researchers interested in adapting other behavioral tasks for online supervised testing (e.g., still face paradigm, book reading tasks, and social touch tasks).

Feasibility and Implementation Strengths

This supervised, synchronous format for online testing proved to be highly feasible, easy to implement, and presented a number of advantages for data collection with infants. Online testing sessions often require fewer research personnel and are shorter in duration compared to laboratory-based testing. For example, online testing eliminates the time required for setting up the physical laboratory space prior to the family's arrival, and critically, the time that is required for infants to get acquainted with the testing environment. In our online study, only one experimenter was required to be online with the family for testing (compared to the addition of a research assistant during in-person testing), and the online session was 30 min in duration compared to 1.5 h when run in the laboratory. Although this discrepancy in duration was partly due to restrictions in what we were able to include in the online assessment (e.g., no inclusion of heart rate measurement during the toy removal task in the online visit), it also reflected a reduction in the time required for infants to become comfortable prior to testing, which tends to be a big source of individual variability in testing times. Online, with families participating from the comfort of their own homes, this "warm up" time is not required, and in the case of our study, the first task typically began within the first 5 min of the session.

The ease of recording the tasks directly through Zoom is another factor that contributed to the high degree of feasibility and easy implementation of this online testing format. In contrast to technological difficulties that may arise when using

video cameras (e.g., uncharged at the time of testing and missing memory cards), recording the tasks through Zoom was highly dependable. When running behavioral tasks, having high quality video recordings are imperative for later analysis. Video recordings are a critical tool in developmental research, as they enable later coding of rich behaviors that may be fleeting in person, they capture the context in which a behavior is embedded (Adolph, 2020), and from an open-science perspective, they allow for widespread data sharing and reproducibility (Gilmore and Adolph, 2017). With the experimenter available to provide live guidance regarding ideal camera angles and lighting, Zoom appears to be a sufficient method for collecting high quality recordings of parent–child interactions. Furthermore, it is a user-friendly technology with which many people are already familiar. Thus, for both research personnel learning to run the online session, and parents participating in the study, there is a minimal learning load from a technology perspective.

We achieved a high rate of task completion for the online study (31/33 to date; 94%, compared to 61/71 for in-person testing; 86%), which may be related to infants' increased comfort in their homes compared to the unfamiliar laboratory environment, reduced overall testing time, and the reliability of Zoom for capturing video recordings.

Challenges and Recommendations

Although the online testing sessions were conducted with a high degree of ease, there are also a number of challenges associated with this format, as well as considerations that will vary depending on the task being adapted. First, not all components of a study will be amenable for remote testing, including the use of specialized technologies like EEG and ECG, which will limit the types of studies researchers can run and the continuity between data collected in person and remotely. Regarding materials, if running a behavioral task is contingent on a specific item (e.g., consistency of the toy across participants is crucial for the validity of the toy removal task), researchers must find a way to mail or drop off materials to families, which may be more or less difficult depending on the location of the research group and other circumstances. Unanticipated issues may arise with the mailing process outside of the researcher's control (e.g., shipping delays, supplier running out of stock, and price increases in the middle of a study). Researchers should have a backup plan for getting any required materials to participants prior to starting data collection. Additionally, in our study, the required material was a fun and exciting toy, which made it appropriate to serve as participant compensation as well. In the case of other tasks where the provided materials would not be well suited to serve as compensation, researchers should consider the added cost of sending materials to families in addition to the funds previously set aside for participant compensation.

Additionally, whereas it is easier to set up multiple camera angles for in-person testing, the reliance on Zoom for all recordings limits the different viewpoints available for recording. For the toy removal task, the single recording is sufficient and

closely resembles the video recordings from in-person testing; however, the free-play task would benefit from an additional “birds-eye” vantage point, which we are able to capture in the laboratory. Different tasks and coding requirements may be more or less amenable to a single viewpoint recording, which should be considered when deciding whether a behavioral task may be appropriate for online adaptation.

Furthermore, information security and participant privacy are important consideration in adapting tasks for online data collection and data storage. Researchers must take precautions to minimize data breaches, which should be coordinated with their respective research ethics board to ensure compliance with institutional guidelines. For example, the use of Zoom as a platform for conducting and recording sessions was approved by our research ethics board as a secure option for collecting data, and recordings were immediately transferred to a secure server for storage. Researchers should also consider whether they can conduct the sessions from a private location when booking sessions (e.g., where others will not be able to see or hear the session) and have the ability to enable a waiting room feature in the video session to ensure unknown persons cannot join the call. These considerations will help ensure participant privacy, confidentiality, and information security.

Testing Environment and Task Setup

Validity

Strengths

Su and Ceci (2021) have highlighted that remote online testing from home includes a trade-off between ecological validity and environmental control, which parallels discussion regarding the tension between “real-world or the lab” testing in psychology more broadly (Hammond and Stewart, 2001; Holleman et al., 2020). In-person home testing has been a cornerstone of developmental research for decades, as measuring infants' real-world behaviors has been highlighted as an important endeavor across developmental fields (e.g., locomotion; Adolph, 2019), and it is thought to be optimized during home-based testing compared to exclusively relying on highly structured, laboratory-based tasks. Furthermore, home-based testing allows for the capture of naturalistic interactions in the settings in which they typically occur, which may afford greater opportunity for measuring family dynamics unaffected by being in a new setting or the presence of other research personnel. Although these benefits of in-person home testing may extend to remote testing from home, environmental control is more difficult when families are tested remotely, as there are likely to be differences in participants' physical living spaces, background noise, and other sources of interference/distraction that are more difficult to minimize when the researcher is not present in the physical space. We argue that in the face of this trade-off, interaction-based tasks that aim to simulate everyday naturalistic interactions between parents and infants are particularly well suited for maintaining their validity during home-based remote testing, *especially* when the format includes live interaction with the researcher. Tight environmental control tends to be less of a concern for interaction-based tasks compared to other forms

of developmental research with infants, such as looking time studies or other visual attention-based paradigms, which are more sensitive to the impact of environmental influences. In a synchronous testing framework, the researcher can maintain the integrity of the study design by ensuring a similar *enough* task setup across participants to provide a sufficient amount of consistency across participants, even in the face of individual differences in families' home environments.

Challenges and Recommendations

Although we believe that interaction-based behavioral tasks are particularly resilient to the lack of tight environmental control obtainable *via* remote online testing, the decision to move to a remote framework may be task dependent. Researchers will need to consider the degree to which completing the task in a naturalistic, yet uncontrolled environment may be an added benefit or detriment to the task validity. For example, in an emotion regulation context, infant attentional strategies serve as an important regulatory strategy (e.g., scanning the room, shifting attention to a novel object, and maintaining gaze on the desired object, such as the toy; Stifter and Braungart, 1995). Scanning patterns may differ depending on the infant's familiarity with their environment (e.g., familiar versus novel room) and the amount of stimulating objects in each environment (e.g., minimalist laboratory testing room compared to a home kitchen full of distractors). Other elements that may introduce a small degree of variability between participants include pets walking into the room during a task, or the noise of other family members in the background. The degree to which these uncontrolled elements impact the validity of a task will depend on the specific behavioral task and serves as an important area of consideration for researchers contemplating moving a task to a remote testing framework. This challenge is similar to what might be encountered with in-person home testing; however, some of these uncontrolled elements may be amplified in a remote framework where the researcher is not on-site to manage some of the environmental differences.

Accessibility, Recruitment, and Racial and Socioeconomic Diversity Strengths

Online testing greatly improves accessibility. Shorter testing sessions and the elimination of travel made possible through online testing offer greater flexibility with respect to scheduling, which is a critical ingredient in mitigating attrition in longitudinal studies. Our laboratory has previously found it difficult to re-recruit infants in the older age range of our longitudinal study (e.g., 29% attrition between 3 to 7 months vs. 51% attrition at 12 months and 57% attrition at 18 months), which is largely due to parents' returning to work and reduced availability. These scheduling constraints are further exacerbated by studies with longer testing sessions. Remote online testing offers greater flexibility for evening testing (e.g., less travel time and sessions are less likely to overlap with infants' bedtimes) and the ability to book back-to-back sessions to accommodate more weekend testing times (e.g., no turnaround time required to clean up

and prepare materials between families), which may facilitate parents' ability to continue their participation in longitudinal studies after returning to work. The elimination of travel, which has been previously identified as a significant barrier to families' participation in developmental research (Sugden et al., 2015), strongly contributes to the accessibility of online testing. The option to participate remotely may increase accessibility for families who live further away from universities, and for families who have moved over the course of a longitudinal study. These benefits are similar to those offered by in-person home testing; however, remote online testing eliminates the need for travel for *both* the family and the researcher, rendering it even more advantageous for flexible scheduling.

The increased accessibility of online testing may also lead to improvements in recruiting more racially and socioeconomically diverse samples (Rhodes et al., 2020; Sheskin et al., 2020; Su and Ceci, 2021). Psychology research has traditionally oversampled from Western, Educated, Industrialized, Rich, Democratic (WEIRD) populations (Henrich et al., 2010), which threatens the generalizability of research findings and further marginalizes low-income and racial minority populations. The elimination of travel may boost participation among families of lower socioeconomic status for whom travel costs may have been a deterrent to participating in laboratory-based testing, and it provides researchers with the option to recruit outside of their direct geographical location. Families who are new to participating in research studies may also feel more comfortable participating from their own homes for the first time (Sheskin et al., 2020).

Challenges and Recommendations

In considering how to maximize a study's accessibility, researchers should try to minimize the materials families require to be eligible for participation. In our study, the only materials required for participation were a laptop or tablet with Zoom capability and a high chair. Families who did not have access to a printer were given the option of providing verbal consent and completing a questionnaire in real-time with the experimenter, such that printing materials beforehand was not a condition for participation. For our free-play task, families were able to use the toys available to them at home, which was fitting for a naturalistic task like this one. Required materials are important for researchers to consider when adapting tasks to increase the accessibility of participation and to consider ways to minimize the burden on participants to source and provide their own materials.

Regarding recruitment, one way that we maximized participation from previously participating families was by expanding the age range at which they were eligible to participate, which allowed us to capture families that had aged out of our more restricted time range. This adjustment was possible for the current tasks because we did not expect significant differences in performance across our expanded age range; however, for other tasks where significant development might be expected within a short window, expanding the age range to maximize participation may not be possible.

Undoubtedly, online testing introduces a new barrier to participation, the requirement of home internet access, which may be disproportionately lacking among low socioeconomic and racial minority populations and may compound issues of “digital divide” across groups (Haight et al., 2014). For example, lower rates of internet access are reported among households with lower incomes, lower levels of education, and recent immigrants (Haight et al., 2014). Recommendations to promote racial diversity in online studies include tailoring recruitment efforts in line with those found to be effective for the specific group of interest (e.g., non-White groups; Sugden and Moulson, 2015), collecting and reporting detailed demographic data, allocating funds for providing participants with mobile hotspots if needed, and exploring the option of mobile testing laboratories when it is safe to implement face-to-face testing (Lourenco and Tasimi, 2020).

Conclusion

Remote online testing is likely to prevail as an enduring method for conducting developmental research beyond the pandemic (Su and Ceci, 2021); thus, generating and evaluating options for conducting studies of varied methodologies and appropriate for different age groups in a remote format are of paramount importance for the field of developmental science. In considering the advantages and disadvantages of the remote testing framework outlined here, we propose that this synchronous format of online testing offers a highly feasible and easy-to-implement option for collecting infant behavioral data remotely, in which the reliability and validity of the task setup and quality of the data are largely preserved. This format offers many of the general benefits of remote unmoderated testing, including greater scheduling flexibility and potential for more diverse samples. Further, the added component of live interaction with the researcher provides additional benefits previously unique to face-to-face testing, such as the ability to ensure a consistent study procedure is followed across participants. We suggest that behavioral interaction-based tasks are particularly amenable to this synchronous testing format, and we encourage the adoption of this framework across other behavioral tasks, beyond the two presented here.

REFERENCES

- Adolph, K. E. (2019). “Ecological validity: mistaking the lab for real life,” in *My Biggest Research Mistake: Adventures and Misadventures in Psychological Research*. ed. R. Sternberg (New York, NY: Sage), 187–190.
- Adolph, K. E. (2020). Oh, behave! *Infancy* 25, 374–392. doi: 10.1111/infa.12336
- Gilmore, R. O., and Adolph, K. E. (2017). Video can make behavioural science more reproducible. *Nat. Hum. Behav.* 1, 1–2. doi: 10.1038/s41562-017-0128
- Haight, M., Quan-Haase, A., and Corbett, B. A. (2014). Revisiting the digital divide in Canada: the impact of demographic factors on access to the internet, level of online activity, and social networking site usage. *Inf. Commun. Soc.* 17, 503–519. doi: 10.1080/1369118X.2014.891633
- Hammond, K. R., and Stewart, T. R. (2001). *The Essential Brunswick: Beginnings, Explications, Applications*. New York, NY: Oxford University Press.
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–83. doi: 10.1017/S0140525X0999152X
- Holleman, G. A., Hooge, I. T., Kemner, C., and Hessels, R. S. (2020). The ‘real-world approach’ and its problems: a critique of the term ecological validity. *Front. Psychol.* 11:721. doi: 10.3389/fpsyg.2020.00721

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, and further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ryerson University Research Ethics Board (REB). Written informed consent to participate in this study was provided by the participants’ legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SS: study conceptualization, data collection, and original draft — writing. MM: study conceptualization, funding acquisition, original draft — feedback and edition. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the Social Sciences and Humanities Research Council of Canada (#435–2017-1438; awarded to MM), the Ontario Ministry of Research, Innovation, and Science (#ER15-11-162; awarded to MM), and Ryerson University.

ACKNOWLEDGMENTS

We thank all the families who participated in this research and who graciously “let us into their homes” to continue our study remotely. We would also like to thank the members of our research team who were instrumental in participant recruitment, including Yasmine Nouredine.

- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Lovibond, P. F., and Lovibond, S. H. (1995). The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the Beck depression and anxiety inventories. *Behav. Res. Ther.* 33, 335–343. doi: 10.1016/0005-7967(94)00075-U
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open. Mind.* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Segal, S. C., Marquis, A. R., and Moulson, M. C. (2021). Are our samples representative? Understanding whether temperament influences infant dropout rates in a longitudinal study at 3 and 7 months. *Infant Behav. Dev.* 65:101630. doi: 10.1016/j.infbeh.2021.101630
- Segal, S. C., and Moulson, M. C. (2020a). What drives the attentional bias for fearful faces? An eye-tracking investigation of 7-month-old infants’ visual scanning patterns. *Infancy* 25, 658–676. doi: 10.1111/infa.12351

- Segal, S. C., and Moulson, M. C. (2020b). Dynamic advances in emotion processing: differential attention towards the critical features of dynamic emotional expressions in 7-month-old infants. *Brain Sci.* 10:585. doi: 10.3390/brainsci10090585
- Sheskin, M., and Keil, F. (2018). The ChildLab.com a video chat platform for developmental research. [Preprint] doi: 10.31234/osf.io/rn7w5
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Stifter, C. A., and Braungart, J. M. (1995). The regulation of negative reactivity in infancy: function and development. *Dev. Psychol.* 31, 448–455. doi: 10.1037/0012-1649.31.3.448
- Su, I., and Ceci, S. (2021). “Zoom Developmentalists”: home-based videoconferencing developmental research during COVID-19. [Preprint] doi: 10.31234/osf.io/nvdy6
- Sugden, N. A., Kusec, A., Meisner, B., and Moulson, M. C. (2015). *Delivering Baby Scientists: Parents’ Perspectives on the Benefits of and Barriers to Participating in Developmental Research*, Philadelphia, USA: Poster presented at the Society for Research on Child Development
- Sugden, N. A., and Moulson, M. C. (2015). Recruitment strategies should not be randomly selected: empirically improving recruitment success and diversity in developmental psychology research. *Front. Psychol.* 6:523. doi: 10.3389/fpsyg.2015.00523
- Tronick, E., Als, H., Adamson, L., Wise, S., and Brazelton, T. B. (1978). The infant’s response to entrapment between contradictory messages in face-to-face interaction. *J. Am. Acad. Child Psychiatry* 17, 1–13. doi: 10.1016/S0002-7138(09)62273-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Segal and Moulson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Baby's Online Live Database: An Open Platform for Developmental Science

Masaharu Kato^{1,2*}, Hirokazu Doi^{2,3}, Xianwei Meng^{2,4}, Taro Murakami^{2,5}, Sachiyo Kajikawa^{2,6}, Takashi Otani^{2,7} and Shoji Itakura^{1,2}

OPEN ACCESS

Edited by:

Sho Tsuji,
The University of Tokyo, Japan

Reviewed by:

Mingdi Xu,
Keio University, Japan
Hisako W. Yamamoto,
Tokyo Woman's Christian University,
Japan

*Correspondence:

Masaharu Kato
maskato@mail.doshisha.ac.jp

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 22 June 2021

Accepted: 21 September 2021

Published: 13 October 2021

Citation:

Kato M, Doi H, Meng X, Murakami T,
Kajikawa S, Otani T and
Itakura S (2021) Baby's Online Live
Database: An Open Platform for
Developmental Science.
Front. Psychol. 12:729302.
doi: 10.3389/fpsyg.2021.729302

¹Center for Baby Science, Doshisha University, Kyoto, Japan, ²Live Database Working Group, Japan Society of Baby Science, Tokyo, Japan, ³School of Science and Engineering, Kokushikan University, Tokyo, Japan, ⁴Graduate School of Human Sciences, Osaka University, Suita, Japan, ⁵Department of Education and Psychology, Kyushu Women's University, Kitakyushu, Japan, ⁶College of Arts and Sciences, Tamagawa University, Machida, Japan, ⁷Department of Psychology, Faculty of Health Science, Kyoto Koka Women's University, Kyoto, Japan

Efficient data collection in developmental studies is facing challenges due to the decreased birth rates in many regions, reproducibility problems in psychology research, and the COVID-19 pandemic. Here, we propose a novel platform for online developmental science research, the Baby's Online Live Database (BOLD), which extends the scope of the accessible participant pool, simplifies its management, and enables participant recruitment for longitudinal studies. Through BOLD, researchers can conduct online recruitment of participants preregistered to BOLD simply by specifying their attributes, such as gender and age, and direct the participants to dedicated webpages for each study. Moreover, BOLD handles participant recruitment and reward payment, thereby freeing researchers from the labor of participant management. BOLD also allows researchers the opportunity to access data that were collected from participants in previous research studies. This enables researchers to carry out longitudinal analyses at a relatively low cost. To make BOLD widely accessible, a consortium was formed within the Japan Society of Baby Science, where members from diverse research groups discussed the blueprint of this system. Once in full-scaled operation, BOLD is expected to serve as a platform for various types of online studies and facilitate international collaboration among developmental scientists in the near future.

Keywords: COVID-19, longitudinal study, reproducibility, developmental science, open science, survey at home, online study

INTRODUCTION

Developmental science investigates the principles of human beings' physical and mental abilities from the perspective of development. In this field of research, babies, children, and their caregivers are recruited for observations, surveys, and experiments.

Currently, developmental science faces three major challenges. First, the population of young people who could be participants in developmental science research is decreasing. The birth rate has decreased in many countries; the average fertility rate of the 37 of organisation for economic cooperation and development (OECD) member countries began declining in 1970 and has been hovering below 2.0 since 1991 (OECD, 2021a). The ratio of people under 15 years old to the total population in OECD countries has diminished from 25.3% in 1980 to 17.7% in 2018 (OECD, 2021b).

The second challenge is reproducibility. In recent years, the standards for publishing experimental research have become stricter in response to the so-called reproducibility problem (Open Science Collaboration, 2015). The reproducibility crisis showed that good, reliable research often requires larger samples. However, researchers in small- and medium-sized laboratories, who do not have sufficient resources, may struggle to achieve this goal. This situation has a particularly negative impact on the career development of newly independent young principal investigators (PIs) and may lead to a shrinking base in developmental science in the future, and ultimately, the decline of the field.

Finally, the ongoing COVID-19 pandemic has highlighted the vulnerability of human experimental research to the unexpected occurrence of public health concerns and natural disasters. Thus, it is desirable to create infrastructure that enables researchers to continue experimental research in an adverse environment.

Creating a platform where experimental research can be conducted online is a promising solution to the issues listed above, as it would allow for research participation without visiting laboratories. Emerging online tools for experimental research, such as the programming libraries of jsPsych (de Leeuw, 2015), PsyToolkit (Stoet, 2010), and Gorilla (Anwyl-Irvine et al., 2020), and cloud-based sourcing platforms such as Amazon Mechanical Turk, have been of great help to researchers when building and deploying their experiments online. Several systems for conducting online experiments have been proposed to use instead of face-to-face experiments (Frank et al., 2017; Scott and Schulz, 2017; Sheskin and Keil, 2018; Mehr et al., 2019; Rhodes et al., 2020). However, while these systems should contribute to solving the challenges noted above, they are insufficient for overcoming these problems, as they do not help to recruit or manage participants, which is what young PIs and researchers in small- and medium-sized laboratories need. In this paper, we propose the Baby's Online Live Database (BOLD), an umbrella database system suitable for participant recruitment and management. Sheskin et al. (2020) have previously emphasized the necessity of such an online platform, and our platform may be the first to be implemented. The main aim of BOLD is to provide a

large-scale (e.g., national) participant database that can be widely used to run experiments in developmental science.

Additionally, BOLD aims to enhance collaborative and longitudinal studies. Using BOLD, researchers can gain access to participants' task history, including detailed information on studies that the participants have completed, and their performance. This allows researchers to both link participants' performance across studies for comparison and perform longitudinal analyses. To achieve this, the participants would need to be engaged in BOLD long-term. Thus, the importance of including research topics that interest and motivate participants to join our database and stay involved is emphasized.

In this perspective paper, we describe the blueprint for BOLD. Implementation is ongoing, and full-scale operation is expected to begin in late summer 2021. Our goal is to make BOLD available to everyone interested in developmental research. Therefore, a working group was formed within the Japan Society of Baby Science (JSBS), in which the core members of the working group, who come from diverse research groups, discuss the basic design of BOLD and how to proceed with it. Once made publicly available, BOLD will drastically reduce the cost of recruiting and managing study participants for researchers. Researchers will be able to reach participants from many districts around Japan, mitigating concerns about selection bias. This will benefit young PIs with limited resources and other researchers who need to lower the cost of conducting developmental research.

BABY'S ONLINE LIVE DATABASE

Below, we describe the grand concept and implementation of BOLD. BOLD is comprised of two main systems: participant management and study management. As the core of BOLD, we adopted the cloud-based research and participant solution system provided by Sona Systems. Sona Systems provides an online/paperless system for participant/research management, and the system has been introduced at over 1,000 universities worldwide. We modified the system's fundamental functions to increase its suitability for developmental science research. The website BOLD, powered by Sona system, can be accessed *via* <https://doshisha-akachan.sona-systems.com/>. The JSBS working group will direct BOLD. JSBS has been in existence for more than 20 years and is financially stable, making it a suitable candidate for sustainable management of BOLD. The Center for Baby Science at Doshisha University will manage the actual administrative work and financial support. We are currently preparing to incorporate the society in the future. Because JSBS is responsible for BOLD, additions or deletions from its working group members does not affect the management of BOLD.

Participant Management System

The participant management system is responsible for managing participant information and reward payments. Caregivers can make an account and register their personal information, including their child's age in months. Once an account has been created, a participant can apply to (or be invited to) studies through

a webpage (or via emails). After participation, reward points that are monetizable are added to the participant's account.

Study Management System

The study management system manages study registration and participant recruitment.

First, a prospective BOLD researcher must apply; then, two or three JSBS working group members will blindly review the application and make a report. Based on the report, the working group will choose the application that they believe will contribute to BOLD's development. After acceptance, researchers can register their studies within BOLD; they must specify the desired attributes of participants (e.g., gender and age), what data they want to collect, and the schedule of data collection. The ethical committee review approval period and number should also be registered. Regarding ethical considerations, researchers must undergo an ethics review at their institution when they plan to use BOLD. At that time, the application for data reuse will be included in advance and will provide a legitimate basis for data sharing. When researchers run an experiment with the assistance of BOLD, they will be asked to commit to making their data available to researchers who have undergone the same admission process.

Based on the registered information, the study management system extracts qualified participants from the participant management system. Information about the study is delivered to the qualified participants on their BOLD page and in their email. The interface for participant recruitment is shown in **Figure 1A**.

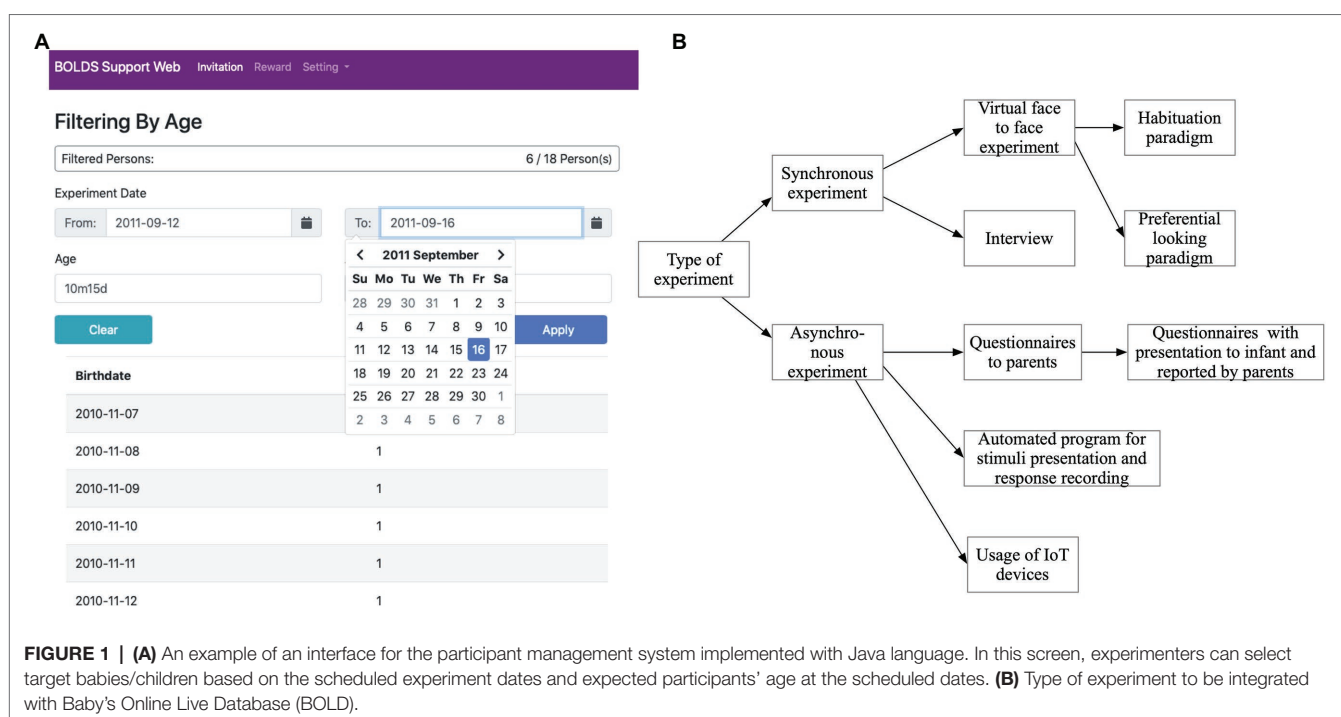
Potential participants receive an invitation email which directs them to the study's page. If they decide to participate, they are prompted to click on the consent button to indicate consent for participation. This can be considered informed consent

and is obtained in a digital form. The FDA allows digital informed consent (United States Food and Drug Administration, 2016). Though initiatives on digital informed consent have begun in Japan (Kogetsu and Kato, 2019), handwritten consent is still customary and therefore is accepted on BOLD. A reminder email will be sent automatically to participants just prior to the scheduled event.

Throughout all stages of data collection, personal information is confined within BOLD. Researchers are provided with only a participant ID, through which they can contact participants. It is possible for researchers to acquire personal information during experiments, such as virtual face-to-face experiments and interviews. However, these are the same as typical, offline experiments. BOLD will not limit the contents of experiments, and researchers must be responsible for the content of their experiments. It is the ethical committee's role to protect participants' personal information collected during experiments.

CONDUCTING STUDIES THROUGH BOLD

An advantage of BOLD is performing longitudinal studies with previously collected data. If previous study participants remain registered in BOLD, researchers can (i) collect new data from the participants and (ii) access participant data from past studies. By linking these datasets, researchers can perform a longitudinal analysis. In addition, basic cognitive, motor, and social developmental performance (the essential dataset) is collected from all participants on enrollment. This will be beneficial in longitudinal studies as the data from the initial time point are already collected.



To collect new data, BOLD navigates potential participants to websites where researchers set up their experiments and surveys. This gives researchers flexibility to conduct various types of studies using any libraries of their choice. The studies conducted are often categorized into two types: synchronous study, where participants and researchers coordinate their time and meet face-to-face over videoconferencing, and asynchronous study, where participants can participate in the survey at a time which is convenient for them, as shown in **Figure 1B**.

In the synchronous study, participants and researchers meet face-to-face *via* a videoconferencing system. Researchers can therefore carry out experiments and surveys, just as in the laboratory. For example, researchers can video-record infants' faces while presenting stimuli and analyze recorded videos offline to quantify a rough estimate of fixation duration. Conventional paradigms of preferential looking and habituation–dishabituation paradigms can also be implemented online.

In the asynchronous study, three main types of studies are feasible. The first is a web-based questionnaire, in which participants answer online questionnaires at a convenient time. Some online survey systems, such as Qualtrics and SurveyMonkey, offer multimedia content presentations. Thus, it is also possible to present movies and collect responses to them. When collecting young children's responses to multimedia content, caretakers can enter information about their children's behavior in the survey form (e.g., Meng et al., 2021). The second method is a pre-programmed study. In this type of study, when a participant accesses a website for research, they are automatically given instructions. Participants follow the instructions and create responses that are stored in the experiment program's server. This type of study is suitable for measuring the behavioral responses of older children and caregivers. It could also be possible to collect eye movement data using libraries such as webgazer.js. However, the validity of web-camera-based eye tracking has only been tested in adult participants, with a few exceptions (e.g., Semmelmann et al., 2017). The third method involves data collection using handy Internet-of-Things (IoT) devices (e.g., the ferro-electret sensor provided by Emfit Ltd. in Finland, used for measuring ballistocardiogram during sleep). The data collected by IoT devices can be transmitted directly to cloud servers and retrieved by researchers. Although the measurement of physiological data using IoT devices remains challenging, the results of such attempts would be significant because raw physiological data contain vast amounts of information that can be analyzed by various methods, per the researchers' choice.

CURRENT STATUS OF BOLD

The implementation of BOLD is still in progress; however, the early registration of participants has already begun. The Center for Baby Science at Doshisha University began registering potential study participants in spring 2020. Registrations were made from different areas in Japan, meaning that researchers can reach people in remote areas and people who are nearby but are not able to travel to the study site due to disability. A trial recruitment period began in May 2020, and registrations rapidly increased,

reaching approximately 400 over 50 days. The key to BOLD's success is creating as large a pool of potential participants as possible. To collect essential minimum data in August 2021, we will soon implement a questionnaire survey for 10-month-old infants on physical and psychological development using the Kinder Infant Development Scale (Hashimoto, 2013). We have set 10 months as the minimum age due to the limitations of our research resources (i.e., we have no experts in early human development as members). We hope to lower this in the future. At present, we would like to begin with typically developing children because diagnostic information about diseases is considered personal information that requires special attention.

PARTICIPANT-ORIENTED PLATFORM

Many caregivers are concerned about whether their parenting style is appropriate. However, we cannot say with confidence that developmental scientists have fully answered their questions and concerns. One way to attract potential participants is making sure that participants' concerns are addressed and their interests are satisfied by joining the platform. It would be effective if participant can make a question to other participants as a participant-driven survey. Another way is by setting up a forum. Many of the concerns caregivers have are individualized and specific and are therefore not likely to be researched. Thus, many caregivers may want to ask caregivers with older children what to do about these problems. This can be achieved by creating a place where participants can raise questions and have them answered. Alternatively, if studies are conducted through BOLD that address caregivers' concerns, it might increase their motivation to register.

Considering this, we carried out a preliminary survey to clarify the topics caregivers are most interested in as reference information for determining the first batch of studies to conduct using BOLD. The total number of participants was 587. The detailed procedure of the preliminary survey and the questionnaire items are described in **Supplementary Material**.

The main results of the first questionnaire block are presented in **Figure 2**. Regarding the most concerning problematic behavior (**Figure 2A**), frequency of choosing the "other" option increases with child's age, indicating diversification of problematic behaviors. **Figure 2B** shows that caregivers have a strong interest in what kind of sports activity is most beneficial for children from the early stages of development. At around 4–5 years old, interest in lessons in "Juku," a private tutoring program for school entrance examination, steeply increases. Among topics related to children's temperament, concerns about shyness and restlessness increase with age (**Figure 2C**).

In the second block of the questionnaire, respondents were asked to choose the most interesting research topics from a list of academic research topics in the developmental science field. The results are summarized in **Supplementary Material**. Broad topics of "Mental Development" and "Brain Development" were most frequently chosen, while more specific topics like "Development of Self Control," "Moral Development," and "Language Development" garnered a relatively small number of votes.

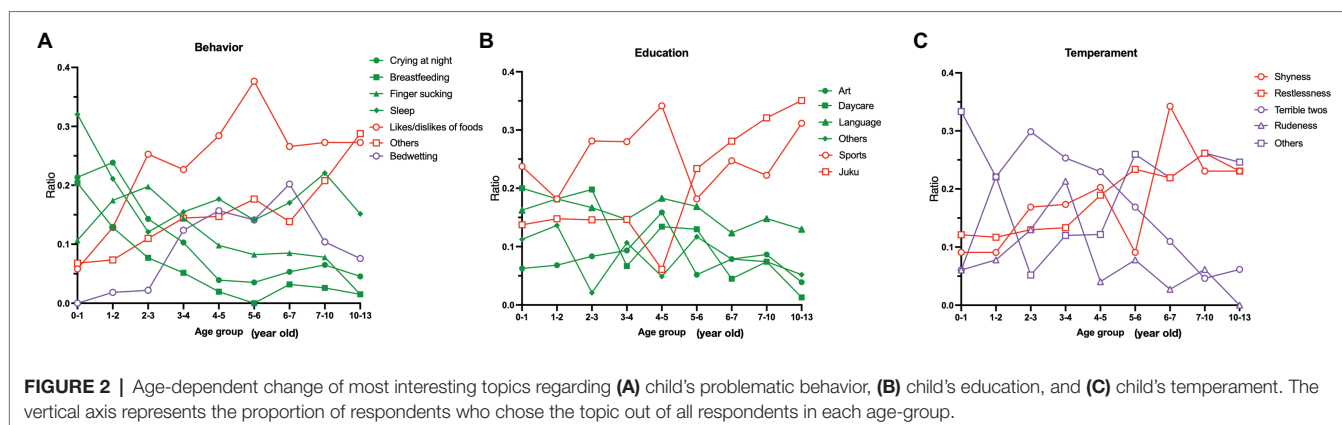


FIGURE 2 | Age-dependent change of most interesting topics regarding (A) child's problematic behavior, (B) child's education, and (C) child's temperament. The vertical axis represents the proportion of respondents who chose the topic out of all respondents in each age-group.

DISCUSSION

Baby's Online Live Database aims to solve problems that developmental scientists are currently facing. This system simplifies participant recruitment and makes it easier to reach a large pool of participants from remote areas of the country and conduct longitudinal studies on human psychological development.

A longitudinal study with a prospective design (e.g., birth cohort study) is a powerful method for understanding the mechanisms of human development. Although there are many birth cohorts (Andersen and Casas, n.d.), our system has two prominent features that make it easier for developmental scientists to conduct longitudinal data analysis. First, our system enables researchers to recruit participants and access data from past studies in which they have participated. This makes it possible for researchers to pursue their interests without needing to obtain the necessary budget to sustain a longitudinal study. Second, new researchers are welcomed to use and make novel contributions to the system. This feature differs from the management of many other birth cohorts, where the chance of joining a longitudinal study and accessing the data are restricted to members of the research groups hosting the cohort. Owing to this openness, BOLD has the potential to accumulate longitudinal data on diverse topics hitherto neglected in existing cohort studies.

The downside of our system is that it is possible for the dataset constructed in our system to become an assortment of independent datasets that are only loosely associated with each other. However, this possibility can be reduced by collecting essential minimum dataset of great interest to many developmental scientists from all participants. We are currently deciding on the types of data to include in the essential minimum dataset, and a physical and psychological development scale (Hashimoto, 2013) should be included in the essential minimum dataset. Performance of popular behavioral tasks, such as delayed gratification tests and preferential looking to social stimuli, is also a good candidate. The inclusion of the essential minimum dataset is beneficial for both researchers and participants. It would be good motivation for researchers to use BOLD if such an attractive dataset was available. The results of the development scale included in the essential minimum dataset would also be interesting to participants.

Sending reports of essential minimum datasets will increase their satisfaction.

Our preliminary questionnaire survey revealed that caregivers' specific concerns about child development change with the child's age. The results showed that caregivers have a strong interest in neurological and psychological development. At the same time, relatively few caregivers chose specific developmental science topics as those of most interest. These results may indicate that caregivers generally have a broad interest in children's psychological development and that their interest is not necessarily restricted to specific cognitive functions. Non-specialists are generally unfamiliar with the recent progress of these research topics and their significance in considering children's development. This would be one reason why these topics, though appealing to developmental psychologists, were not popular among caregivers, thus representing the gap between caregivers' and researchers' interests. Bridging this gap may make caregivers more willing to participate in the researcher's study. To achieve this, researchers should increase awareness among caregivers regarding the importance of the research topics that seem at first glance irrelevant to their children's development and, thus, uninteresting. Alternatively, BOLD may conduct a survey according to the caregivers' interest and send them reports of the results. This gesture will make the caregivers aware that members of BOLD do care about what they truly want to know. After such experiences, caregivers may agree to participate in other more researcher-oriented studies.

Baby's Online Live Database aims to provide solutions to the problems that developmental scientists are currently facing, primarily by reducing the cost of participant recruitment and management and simplifying the process of conducting longitudinal analyses. As the number of users increases, BOLD will become a research platform beneficial for researchers as well as participants and caregivers. It is still a small initiative, but we welcome collaborators to make it a large and international system in the future.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

MK conceived the idea. MK and HD conceptualized the idea and wrote the original draft of the manuscript. MK, HD, XM, TM, SK, TO, and SI worked on implementing the platform.

REFERENCES

- Andersen, A.-M. N., and Casas, M. (n.d.). Birthcohorts.net. Available at: <https://www.birthcohorts.net> (Accessed March 1, 2021).
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: an online behavioral experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x
- de Leeuw, J. R. (2015). jsPsych: a JavaScript library for creating behavioral experiments in a web browser. *Behav. Res. Methods* 47, 1–12. doi: 10.3758/s13428-014-0458-y
- Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., et al. (2017). A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy* 22, 421–435. doi: 10.1111/inf.12182
- Hashimoto, K. (2013). Validity of the family-rated Kinder Infant Development Scale (KIDS) for children. *Pediatr. Ther.* 3:153. doi: 10.4172/2161-0665.1000153
- Kogetsu, A., and Kato, K. (2019). Notes on the use of electronic methods for research participation (in Japanese). Report at Japan agency for medical research and development. Available at: <https://www.amed.go.jp/content/000047937.pdf> (Accessed September 15, 2021).
- Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., et al. (2019). Universality and diversity in human song. *Science* 366:eaax0868. doi: 10.1126/science.aax0868
- Meng, X., Kato, M., and Itakura, S. (2021). Development of synchrony-dominant expectations in observers. *Soc. Dev.* 1–13. doi: 10.1111/sode.12556
- OECD (2021a). Fertility rates (indicator). doi: 10.1787/8272fb01-en
- OECD (2021b). Young population (indicator). doi: 10.1787/3d774f19-en
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science* 349:aac4716. doi: 10.1126/science.aac4716
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002

All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by MEXT Promotion of Distinctive Joint Research Center Program (Grant Number: JPMXP0619217850) implemented at Doshisha University Center for Baby Science.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.729302/full#supplementary-material>

- Semmelmann, K., Hönekopp, A., and Weigelt, S. (2017). Looking tasks online: utilizing webcams to collect video data from home. *Front. Psychol.* 8:1582. doi: 10.3389/fpsyg.2017.01582
- Sheskin, M., and Keil, F. (2018). TheChildLab.com: a video chat platform for developmental research. *PsyArXiv [Preprint]*. doi: 10.31234/osf.io/rn7w5
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Stoet, G. (2010). PsyToolkit: a software package for programming psychological experiments using Linux. *Behav. Res. Methods* 42, 1096–1104. doi: 10.3758/BRM.42.4.1096
- United States Food and Drug Administration (2016). Use of electronic informed consent in clinical investigations – questions and answers guidance for institutional review boards, investigators, and sponsors. FDA-2015-D-0390.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Kato, Doi, Meng, Murakami, Kajikawa, Otani and Itakura. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Online Testing Yields the Same Results as Lab Testing: A Validation Study With the False Belief Task

Lydia Paulin Schidelko^{*†}, Britta Schünemann[†], Hannes Rakoczy and Marina Proft

Department of Developmental Psychology, University of Göttingen, Göttingen, Germany

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Lindsay Bowman,
University of California, Davis,
United States
Rose M. Scott,
University of California, Merced,
United States

*Correspondence:

Lydia Paulin Schidelko
lydiapaulin.schidelko@
uni-goettingen.de

[†]These authors share first authorship

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 April 2021

Accepted: 16 September 2021

Published: 13 October 2021

Citation:

Schidelko LP, Schünemann B,
Rakoczy H and Proft M (2021) Online
Testing Yields the Same Results as
Lab Testing: A Validation Study With
the False Belief Task.
Front. Psychol. 12:703238.
doi: 10.3389/fpsyg.2021.703238

Recently, online testing has become an increasingly important instrument in developmental research, in particular since the COVID-19 pandemic made in-lab testing impossible. However, online testing comes with two substantial challenges. First, it is unclear how valid results of online studies really are. Second, implementing online studies can be costly and/or require profound coding skills. This article addresses the validity of an online testing approach that is low-cost and easy to implement: The experimenter shares test materials such as videos or presentations via video chat and interactively moderates the test session. To validate this approach, we compared children's performance on a well-established task, the change-of-location false belief task, in an in-lab and online test setting. In two studies, 3- and 4-year-old received online implementations of the false belief version (Study 1) and the false and true belief version of the task (Study 2). Children's performance in these online studies was compared to data of matching tasks collected in the context of in-lab studies. Results revealed that the typical developmental pattern of performance in these tasks found in in-lab studies could be replicated with the novel online test procedure. These results suggest that the proposed method, which is both low-cost and easy to implement, provides a valid alternative to classical in-person test settings.

Keywords: online studies, validation study, developmental psychology, psychology methods, Theory of Mind, false belief

INTRODUCTION

Developmental research largely depends on collecting data from children. While varying in methods, set-ups and concrete testing sites, so far, most research has been conducted in an interpersonal, face-to-face setting between an experimenter and a child. Thus, with the beginning of the COVID-19 pandemic, most well-established testing routines were suddenly disrupted and the need for new, safe, and contact-free ways to test children for developmental studies arose.

In the last decade, online testing for psychological research already became more and more prominent for adult studies, with several thousand participants taking part in social science experiments every day on platforms like *Amazon Mechanical Turk (MTurk)* and *Prolific* (Bohannon, 2016). More recently, developmental researchers have started to establish first online platforms for children, including *Lookit* (Scott and Schulz, 2017; Scott et al., 2017) and *Discoveries Online* (Rhodes et al., 2020), that both use an unmoderated set-up (where children and families do not interact with the researchers), and *TheChildLab.com* (Sheskin and Keil, 2018) that uses

a moderated set-up (where the experimenter calls the families via video chat). However, existing platforms and paradigms are not always available for everyone, because of high costs (e.g., for the experimental testing software), programming requirements (e.g., JavaScript), mandatory software downloads or data protection regulations of the software that do not align with the policies of the research institution. Against this background, when we had to close our lab in March 2020, we decided to establish our own moderated testing paradigm for children. In this article, we want to present this novel set-up and validate it as a suitable, safe and broadly accessible tool for online data collection with children.

In our paradigm, we video call families via the software *BigBlueButton* (BBB) and the experimenter then interacts with the children with the help of animated videos or slides. The combination of BBB and screen-sharing comes with several advantages. Concerning the software, BBB is a free, open source, on-premises software. Additionally, once it is established, it comes with low technical requirements both on side of the experimenter as well as the participant as it runs in all common browsers. Furthermore, the servers for BBB are hosted locally, in our case in our institute. Thus, the use of this software allows researchers to adhere to the highest data protection standards, since only the host can access usage and meta-data. Note, however, that while using BBB offers clear advantages, our general set-up is not limited to BBB but is in principle applicable to almost every video chat software that allows screen sharing.

Having set up a technically suitable paradigm, the most pressing question concerns the data quality that can be obtained by testing children with it. Is our moderated online paradigm really appropriate for (remote) data collection? To answer these questions, we wanted to validate our method. Specifically, we tested whether we can conceptually replicate the effects found in in-lab face-to-face settings in analogous studies implemented in our new online paradigm. Importantly, to avoid population-based effects that could explain potential differences between online and in-lab testing, we drew the samples for both paradigms from the same population: our database of parents who had previously given consent to participate with their children in developmental studies. Both samples were thus comparable concerning (a) socio-demographic variables (age and gender were measured, but the sample is also likely to be comparable concerning other socio-demographic variables, e.g., living environment, as the database only includes families living in and around the same city), (b) familiarization with developmental studies (86% of the children participating in an online study participated in at least one other in-person study in our lab before), and (c) incentive structure (we did not directly compensate parents or children for either paradigm).

For the comparison of the two methods, we used a well-established social-cognitive task: the standard false belief (FB) task (Wimmer and Perner, 1983). The FB task is designed to tap children's ability to attribute subjective mental states to others and is generally seen as the litmus test for having a Theory of Mind (ToM). In its standard version, children see a vignette (acted out with puppets) in which an agent puts an object in one of two boxes and leaves the scene. In her absence the object is transferred to the other box and children are then asked to

predict where the agent will look for her object upon her return. Results from countless live studies show that children typically start to master this verbal version of the FB task around the age of four, with younger children falsely predicting that the agent will look for her object where it really is (see Wellman et al., 2001). In addition, we administered the structurally analogous true belief (TB) version of the task. Originally designed to control for extraneous task demands in the FB version, recent studies reported a paradoxical picture: once children master the FB task, they begin to fail the TB task. The TB and FB tasks are thus highly negatively correlated between 3 and 5 such that children first pass the TB and fail the FB task and then show the reverse pattern (see Fabricius et al., 2010; Perner et al., 2015; Oktay-Gür and Rakoczy, 2017). This strange effect in the TB task does not seem to document a conceptual limitation, though. One possibility is that it rather reflects children's sensitivity to task pragmatics that they develop on the basis of their growing Theory of Mind. Several studies reveal that the more advanced in ToM children are, the more pragmatically sensitive they become, and the more they get confused by the triviality of the TB test question given the shared perspective of the experimenter and the child ("Why is the experimenter asking me such a stupid question? I guess there must be a more complex answer than the obvious one"; see Oktay-Gür and Rakoczy, 2017; Rakoczy and Oktay-Gür, 2020). In line with the idea of a high pragmatic component of the TB effect, once the task is modified to become less pragmatically confusing (either by converting it into non-verbal format, or by changing the context so that the question now is less trivial) the effect goes away and children perform competently from age 3 onward without any decline in performance. This is highly relevant for present purposes as it shows that the TB test question in its standard version seem to present a very sensitive measure of children's susceptibility to task pragmatics. The TB task therefore lends itself perfectly as a very stringent test for the comparability of live vs. online testing in even subtle respects of verbal interaction and interpretation.

To validate our online set-up, we thus compare children's performance in the two testing formats (in-lab and online): Do the two paradigms lead to comparable results? This question is not trivial. In fact, the existing literature suggests that there are several indicators that (moderated) online testing might indeed lead to different results. On a general level, there is the video deficit effect (VDE): the phenomenon that children solve the same task later and less accurately when the task is presented in a video than when it is presented by a person (Anderson and Pempek, 2005). The VDE has been found for a variety of tasks such as word-learning (e.g., Thierry and Spence, 2004), object-retrieval (e.g., Troseth and DeLoache, 1998) and imitation (e.g., Klein et al., 2006). Recently, it has also been documented for the FB task: 4- and 5-year-old (who usually pass the task) failed to correctly predict the agent's behavior when the story was presented on a video (Reiß et al., 2014, 2019). Furthermore, there is first data from TheChildLab.com concerning moderated online testing more specifically (Sheskin and Keil, 2018). While in general children tested online provided expected answers on a variety of classical tasks, the FB task seemed to be especially hard: Only 9- to 10-year-olds reliably solved the task, while the 5- to

8-year-olds performed at chance level, opening a gap of around 4–5 years compared to standard in-lab testing results.

For the present validation project, we thus collected data on 3- to 4-year-old children's FB and TB understanding in two online studies and compared it to data we obtained from previous in-lab studies with closely matched protocols. Data from in-lab testing was collected pre-COVID and (partly) reanalyzed for the purpose of the current study (for more details, see **Supplementary Material**). In Study 1, we compared children's performance on the standard FB task between the in-lab and online test setting. In Study 2, we widened the focus and tested whether children's more complex performance patterns in TB and FB tasks would differ between in-lab and online test setting.

STUDY 1

Methods

Participants

The final sample includes 112 monolingual German speaking children aged 36–58 months (mean age = 44.28 months; 56 girls; 64 of them tested in an online test session [mean age = 44.22 months; 31 girls (48%)]); 48 [mean age = 44.35 months; 25 girls (52%)]¹ in an in-lab test session). Mean age did not differ between settings [$t(110) = 0.120$, $p = 0.905$]. All children live in and around the same medium sized German university town, that is generally characterized by mixed socio-economic backgrounds². Six additional children were tested but not included in data analyses because of uncooperative behavior (online setting: $n = 3$), technical issues during the test session (online setting: $n = 1$), parental interference during the test session (online setting: $n = 1$) and language issues³ (in-lab setting: $n = 1$). Children in this and the subsequent study were recruited from a databank of children whose parents had previously given consent to experimental participation.

Design

All children received two trials of a standard change-of-location FB task. The order and direction of location change (from left to right or vice versa) of the trials were counterbalanced. The tasks were presented either as videos in an online testing format or acted-out in an in-lab setting (for comparable scripts and stimuli and a detailed overview of how the online and in-lab tasks were implemented, see **Supplementary Material**).

Materials and Procedure

False belief task

In the FB task (Wimmer and Perner, 1983), Protagonist A (for example, the boy) placed his object (for example, his ball) in

one of two boxes (box 1). In his absence, protagonist B (for example, the girl) moved the ball to the other box (box 2) and the experimenter (E) asked the test question “When the boy returns, where will he look for the ball first?” (Correct answer: box 1) (For additional control questions, see **Supplementary Material**).

Set-Up

Moderated online study

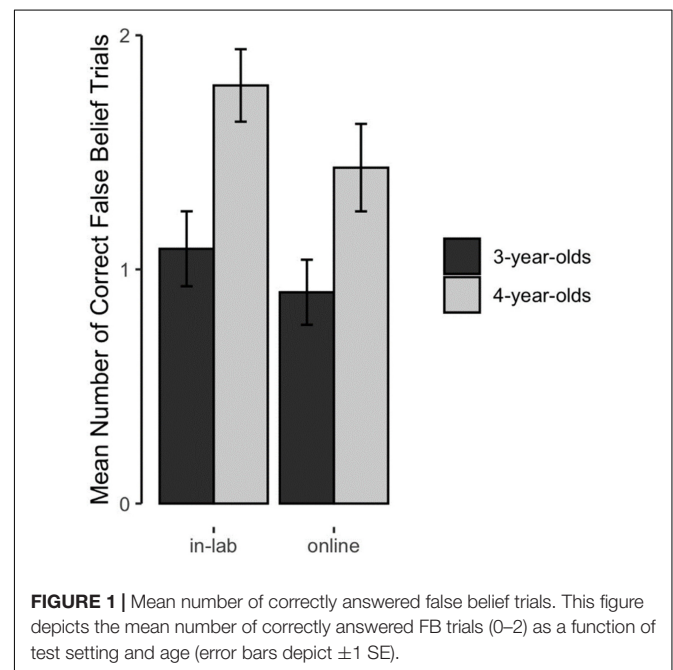
In the online test setting, one female experimenter (E) presented the tasks remotely (on a computer screen, no smartphone) via a video conferencing platform (mainly BigBlueButton, in case of technical issues: Zoom). During the test session, the child and E communicated via audio and video streaming. The story lines of the FB task were visually implemented as short video clips (created with the animation software Vyond™ © 2021 GoAnimate). The child watched the video clips while E told the story lines. At the end of each story line, E asked the control and test questions.

In-lab study

In the in-lab setting, children were tested in single sessions by two female experimenters in the laboratory. E1 first acted out the FB task with little figures and then asked the control and test questions.

Results and Discussion

Figure 1 shows children's performance on the FB test question as a function of age and test setting. In accordance with the literature, we would expect that children's performance on the standard FB task increases with age. If the test setting has an impact on children's performance in this task, there should be a difference between settings most likely in that the effect of age on children's performance should be different between settings.



¹The original sample from Schünemann et al. (2022) included sixty-one 2 1/2–4 1/2-year-old children. For the purpose of the current study, we reduced the data set to a relevant subset of children between 3 and 4 1/2 years (for more details, see **Supplementary Material**).

²Note that we did not collect any data on race, educational level or socio-economic background.

³For the original purpose of the lab study, children were required to be monolingual German (see Schünemann et al., 2022).

For this reason, we set up a Generalized Linear Mixed Model with binomial error structure and a logit link function. As dependent variable, we included children's success on each test trial. To test for an effect of setting and whether the effect of age on performance is different between settings, we included test setting and age measured in months⁴ and their interaction. To account for repeated measures, we included children's ID as random intercept effect. We checked for the model's stability by calculating estimates after case wise exclusion of participants. This revealed a stable model. We also checked for multicollinearity (all $VIFs \leq 1.001$).

We compared this full model to a null model which included age and the random intercept. This comparison was not significant (likelihood ratio test: $\chi^2 = 0.509$, $df = 2$, $p = 0.775$). Likewise, a closer look at the model revealed that the interaction effect of test setting and age was not significant ($b = -0.840$, $p = 0.567$). Also, the main effect for test setting was not significant ($b = -0.297$, $p = 0.813$). Only the main effect for age was significant ($b = 3.789$, $p = 0.013$). Thus, in accordance with the literature, children's performance increased with age. However, in which setting, in-lab or online, the study was conducted did not impact children's performance.

STUDY 2

Methods

Participants

Seventy-six 36- to 53-month-old native German speaking children were included in the final sample (mean age = 43.76 months; 38 girls). Forty-nine children were tested in an online test setting [mean age = 43.49 months; 23 girls (46%)]. Twenty-seven [mean age = 44.26 months, 15 girls (56%)]⁵ were tested in an in-lab test setting. Mean age did not differ between settings [$t(74) = 0.605$, $p = 0.547$]. The children live in and around the same medium sized German university town, that is generally characterized by mixed socio-economic backgrounds⁶. Five additional children were tested in the online test setting but excluded from analysis because they were uncooperative ($n = 4$) or had severe language issues (e.g., could not follow the story line and the experimenter's questions; $n = 1$).

Design

Children again received two trials of a standard change-of-location FB task. Additionally, they received two trials of the TB condition. The two trials of a condition (FB or TB) were presented in blocks. The order of the two blocks and sides of the two trials within the blocks were counterbalanced. The tasks were presented either as an animated slide show in an online testing format or acted-out in an in-lab setting (for comparable scripts and a detailed overview of how the online and in-lab tasks were implemented, see **Supplementary Material**).

⁴Age was z-standardized.

⁵The original sample from Oktay-Gür and Rakoczy (2017, *Exp. 2*) included 171 participants. For the purpose of the current study, we reduced the data set to the relevant subset (for more details, see **Supplementary Material**).

⁶Note that we did not collect any data on race, educational level, or socio-economic background.

Material and Procedure

False belief and true belief task

The protocol was slightly adapted from the classic change-of-location task by Wimmer and Perner (1983) used in Study 1 in that E placed the object in the box and moved the object from the first to the second location in the protagonist's absence (FB) or after her return (TB). After that (TB) or after the protagonist's return (FB), E asked the test question "Where does the protagonist think that the toy car is?" [Correct answer: box 1 (FB), box 2 (TB)] (For additional control questions, see **Supplementary Material**).

Set-Up

Moderated online study

The same set-up was used as in Study 1. The tasks were presented in a slide show, which was displayed on the child's screen via the platform's screen sharing function. While the child was watching the animated slide show, E told the child the story line and asked the control and test questions.

In-lab study

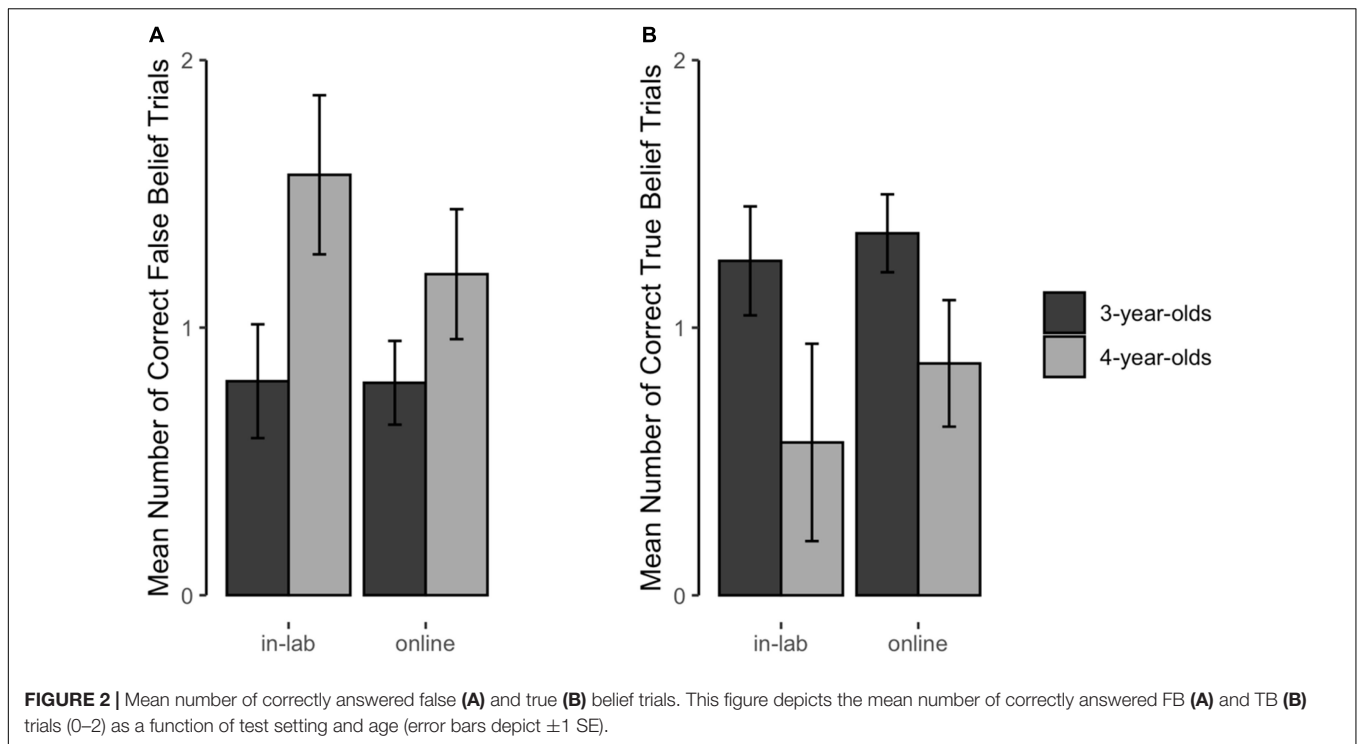
In the in-lab format, children were tested as in Study 1 in single sessions by one of five experimenters in the laboratory or in a quiet room of children's day care.

Results and Discussion

Figure 2 shows children's performance on the FB (a) and TB (b) test questions as a function of age and test setting. In accordance with the literature, we would expect an interaction between age and the belief type: Children performance on the FB task increases with age while it decreases for the TB task. If the test setting has an impact, this interaction of age and belief type should be different between settings.

Again, we set up a Generalized Linear Mixed Model with binomial error structure and a logit link function and success on test trial as dependent variable. To test for the effect of test setting on the interaction of age and belief type, we included test setting, age and belief type and their interactions in the model. To account for repeated measures, we included children's ID as random intercept effect. The model was stable and not multicollinear (all $VIFs = 1$). Again, we compared this full model to a null model. The null model included age, belief type, their interaction and the random intercept.

This full-null model comparison was not significant (likelihood ratio test: $\chi^2 = 2.312$, $df = 4$, $p = 0.679$). Likewise, a closer look at the model revealed neither a significant 3-way-interaction of test setting, age and belief type ($b = 0.616$, $p = 0.268$), nor any interaction with test setting (with age: $b = -0.196$, $p = 0.606$; with belief type: $b = 0.453$, $p = 0.376$). Also, there was no main effect for test setting ($b = -0.177$, $p = 0.617$). In contrast, the interaction effect of age and belief type was significant ($b = -1.656$, $p < 0.001$). Thus, in accordance with the literature, children's performance increased with age for the FB task and decreased for the TB task. The test setting did not have an impact.



GENERAL DISCUSSION

Here, we present and validate a new moderated online testing paradigm for developmental studies. In this paradigm we call families via the video chat software BigBlueButton where the experimenter then interacts with the child with the help of animated videos or slides. The main question regarding the validity of this paradigm was whether it yields results comparable to and converging with in-lab methods. To address this question, we directly compared children's performance in this online paradigm with data from pre-COVID in-lab testing in a standard false belief (FB) and matching true belief (TB) task (Wimmer and Perner, 1983; Oktay-Gür and Rakoczy, 2017). Importantly, we drew samples for both methods from the same database. Thus, all participants were drawn from one population and live in the same local environment. Moreover, in-lab and online samples were matched for age and gender. This reduced potential population-based effects and allowed us to compare the two methods in a very direct and stringent way.

We found no differences between the two testing formats. First, in both studies, 3- and 4-year-olds' performance in the online FB task was equivalent to their performance in the acted-out in-lab versions of the task as well as to what we would expect in that age range given the widely documented "4-year-revolution" of mastering standard FB tasks (Perner, 1991; Wellman et al., 2001). Second, in accordance with previous studies, we found a characteristic performance pattern in FB and TB tasks such that children with age become more proficient in the former while becoming less proficient in the latter. This pattern held equally in both testing formats, with no difference between the in-lab and online tests.

By using our moderated online testing paradigm, we thus replicated children's performance from in-lab testing in samples that were drawn from the same population and without facing issues of data loss. Crucially, however, our paradigm does not only seem to closely match interpersonal, face-to-face testing in terms of "cold" indicators such as data quality. Moreover, it also seems to resemble live set-ups in terms of the naturalness and pragmatics of the interaction: when asked a trivial test question (about an agent's true belief), children showed the same response patterns in the online and the live version. One possible interpretation based on recent research (Rakoczy and Oktay-Gür, 2020) is that children were equally prone to draw pragmatic inferences based on their shared perspective with the experimenter, and fall prey to pragmatic confusions in the online setting as in the in-lab setting. In conclusion, our method seems to be a valid and promising instrument for developmental research.

At the same time, the present results leave open many crucial questions. First, in contrast to previous work (e.g., Anderson and Pempek, 2005; Reiß et al., 2014), we found no indication of the video deficit effect (VDE). Thus, watching video presentations (as children did in our Study 1) does not always seem to disrupt children's performance in comparison to live demonstrations. But why didn't the well-documented VDE occur in our paradigm? What is the crucial difference between the cases in which a VDE occurs (in many previous studies) and cases in which it does not (like the present one)? So far, we can only speculate. One crucial difference between the online format and "classical" video presentations is that in our online paradigm the video and the experimenter are both on the screen while in the classic version only the video is presented on-screen with

the experimenter sitting next to the child as a live interaction partner who asks test questions. Thus, while in the classic format the child has to handle two parallel worlds (on-screen and live), in the online version all relevant information is presented on-screen, potentially helping the child to encode the video more easily. Other potential influencing factors might be related to the sample (including children's age and their drastically increased familiarity with media use during the pandemic) or the specific task type (see Strouse and Samson, 2021). More future research is needed to systematically test the different conditions under which the VDE occurs in relation to online research.

Second, again in contrast to previous work (Sheskin and Keil, 2018), we found no difference in children's relative performance on belief tasks between online and in-lab settings. Thus, administering the task in a moderated testing paradigm *per se* does not seem to negatively influence children's performance. But then, why were there these gaps in previous work? What is the difference between those cases in which online testing is detrimental to performance and those, like the present one, in which it is not? Again, so far, we can only speculate. When we compare our studies to previous ones, at least two differences emerge: Sheskin and Keil (2018) only presented color coded pictures to the children, whereas we implemented a step-by-step analogous video (or animated slide show) version of the acted-out task version using carefully designed online stimuli [e.g., an animated human hand acting out the change of location and (pre-recorded) verbal interaction between protagonists in the story line onscreen and the experimenter; for more details on scripts and stimuli, see **Supplementary Material**]. This suggests that subtle details of online implementations might matter. Another crucial difference is that we had the opportunity to directly compare the data we obtained from the two methods (online and live) rather than loosely contrasting online data to effects from the literature. For this direct comparison we drew the samples from the same population, while previous studies mostly document a more diverse, broader distributed sample in their online compared to in-lab studies (see also Rhodes et al., 2020). Given these differences, it seems plausible to assume that previous work might have underestimated children's performance in (moderated) online paradigms due to population-based effects. Note, however, that although our samples were drawn from the same population, we cannot exclude selective processes in our studies either. There might be a some sort of selection regarding which parents of our population agreed to online testing. Such processes might have led to a less diverse sample and an overestimation of children's performance. Future research is needed to address this possibility and systematically test for the effects of different population-specific parameters such as socio-economic status, living environment, mobility, closeness to the research institute or time flexibility. For example, even though samples for online and in-lab studies are drawn from the same general population (database), do the sub-samples that respond to live vs. online study invitations differ in subtle demographic respects? Also, note that absence of evidence for a difference between in-lab and online testing of belief tasks, of course, does not amount to evidence of absence of any such potential differences. Future research needs to investigate more

systematically and stringently whether there is really no effect of test setting. Such an approach of Bayesian null hypothesis testing will require a larger sample than the current one and will be possible once children can be tested again in an in-lab setting.

The overarching aim of the present project was to find a method for testing children online that is secure, low-cost and easy to implement while yielding comparable results to interpersonal, face-to-face in-lab testing. The results of our two validation studies suggest that with our moderated online testing paradigm we successfully designed such a tool. Future work should now focus on developing the tool further, especially testing its suitability concerning different task types (e.g., more interactive ones that require spontaneous interventions by the child) or different dependent variables (e.g., pointing or eye-tracking). Hopefully, the broader implementation and development of this paradigm then paves the way for more online research in the future, as it has the potential to make developmental research more accessible to a wider audience of participants and researchers.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article are available in the **Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the Local Legislation and Institutional Requirements. Written informed consent for participation was not provided by the participants' legal guardians/next of kin because parts of the studies were conducted online. In the online studies, parents/legal guardians gave verbal consent before the testing was started. Verbal consent was recorded and stored separately from the recording of the test session. For the studies conducted in the laboratory or day care, parents/legal guardians gave written consent.

AUTHOR CONTRIBUTIONS

MP, LS, BS, and HR contributed to conception and design of the study. LS did part of the data collection. BS performed the statistical analysis. MP, LS, and BS wrote the sections and first draft of the manuscript. HR supervised the planning and execution process, provided resources for the data collection, and gave critical review and commentary on the draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 254142454/GRK 2070, Evangelisches Studienwerk Villigst, and Studienstiftung des Deutschen Volkes.

ACKNOWLEDGMENTS

We thank Jana Rechenburg and Anna Lueb for help with data collection, and Marlen Kaufmann for the organization of data collection.

REFERENCES

- Anderson, D. R., and Pempek, T. A. (2005). Television and very young children. *Am. Behav. Sci.* 48, 505–522. doi: 10.1177/0002764204271506
- Bohannon, J. (2016). Mechanical Turk upends social sciences. *Science* 352, 1263–1264. doi: 10.1126/science.352.6291.1263
- Fabricius, W. V., Boyer, T. W., Weimer, A. A., and Carroll, K. (2010). True or false: Do 5-year-olds understand belief? *Dev. Psychol.* 46, 1402–1416. doi: 10.1037/a0017648
- Klein, A., Hauf, P., and Aschersleben, G. (2006). The role of action effects in 12-month-olds' action control: A comparison of televised model and live model. *Infant Behav. Dev.* 29, 535–544. doi: 10.1016/j.infbeh.2006.07.001
- Oktay-Gür, N., and Rakoczy, H. (2017). Children's difficulty with true belief tasks: competence deficit or performance problem? *Cognition* 162, 28–41. doi: 10.1016/j.cognition.2017.05.002
- Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: MIT Press.
- Perner, J., Huemer, M., and Leahy, B. (2015). Mental files and belief: A cognitive theory of how children represent belief and its intensionality. *Cognition* 145, 77–88. doi: 10.1016/j.cognition.2015.08.006
- Rakoczy, H., and Oktay-Gür, N. (2020). Why Do Young Children Look so Smart and Older Children Look so Dumb on True Belief Control Tasks? An Investigation of Pragmatic Performance Factors. *J. Cogn. Dev.* 1, 1–27. doi: 10.1080/15248372.2019.1709467
- Reiß, M., Becker, A., and Krist, H. (2014). Gibt es einen Videodefiziteffekt bei Aufgaben zur Theory of mind? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie* 46, 155–163. doi: 10.1026/0049-8637/a000112
- Reiß, M., Krüger, M., and Krist, H. (2019). Theory of Mind and the Video Deficit Effect: Video Presentation Impairs Children's Encoding and Understanding of False Belief. *Media Psychol.* 22, 23–38. doi: 10.1080/15213269.2017.1412321
- Rhodes, M., Rizzo, M. T., Foser-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing Developmental Science via Unmoderated Remote Research with Children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Schünemann, B., Schidelko, L. P., Proft, M., and Rakoczy, H. (2022). Children understand subjective (undesirable) desires before they understand subjective (false) beliefs. *J. Exp. Child Psychol.* 213:105268. doi: 10.1016/j.jecp.2021.105268
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (Part 2): Assessing the Viability of Online Developmental Research, Results From Three Case Studies. *Open Mind* 2017:1. doi: 10.1162/OPMI_a_00001
- Scott, K. M., and Schulz, L. E. (2017). Lookit: A new online platform for developmental research. *Open Mind* 2017:2. doi: 10.1162/OPMI_a_00002
- Sheskin, M., and Keil, F. (2018). TheChildLab.com: a video chat platform for developmental research. *PsyArXiv*. 30:2018. doi: 10.31234/osf.io/rn7w5
- Strouse, G. A., and Samson, J. E. (2021). Learning From Video: A Meta-Analysis of the Video Deficit in Children Ages 0 to 6 Years. *Child Dev.* 92, e20–e38. doi: 10.1111/cdev.13429
- Thierry, K. L., and Spence, M. J. (2004). A real-life event enhances the accuracy of preschoolers' recall. *Appl. Cogn. Psychol.* 18, 297–309. doi: 10.1002/acp.965
- Troseth, G. L., and DeLoache, J. S. (1998). The medium can obscure the message: Young children's understanding of video. *Child Dev.* 69, 950–965. doi: 10.1111/j.1467-8624.1998.tb06153.x
- Wellman, H., Cross, D., and Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Dev.* 72, 655–684. doi: 10.1111/1467-8624.00304
- Wimmer, H., and Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13, 103–128. doi: 10.1016/0010-0277(83)90004-5

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.703238/full#supplementary-material>

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Schidelko, Schünemann, Rakoczy and Proft. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Tale of Three Platforms: Investigating Preschoolers' Second-Order Inferences Using In-Person, Zoom, and Lookit Methodologies

Elizabeth Lapidow^{1*}, Tushita Tandon¹, Mariel Goddu² and Caren M. Walker¹

¹Department of Psychology, University of California, San Diego, San Diego, CA, United States, ²Department of Psychology, Harvard University, Cambridge, MA, United States

OPEN ACCESS

Edited by:

Sho Tsuji,
The University of Tokyo, Japan

Reviewed by:

Valerie Kuhlmeier,
Queen's University, Canada
Bruce Hood,
University of Bristol, United Kingdom

*Correspondence:

Elizabeth Lapidow
elapidow@ucsd.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 27 June 2021

Accepted: 09 September 2021

Published: 13 October 2021

Citation:

Lapidow E, Tandon T, Goddu M and
Walker CM (2021) A Tale of Three
Platforms: Investigating Preschoolers'
Second-Order Inferences Using
In-Person, Zoom, and Lookit
Methodologies.
Front. Psychol. 12:731404.
doi: 10.3389/fpsyg.2021.731404

As a result of the COVID-19 pandemic, online methodologies for developmental research have become an essential norm. Already, there are numerous options for recruiting and testing developmental participants, and they differ from each other in a variety of ways. While recent research has discussed the potential benefits and practical trade-offs of these different platforms, the potential empirical consequences of choosing among them are still unknown. It is critical for the field to understand not only how children's performance in an online context compares to traditional settings, but also how it differs *across* online platforms. This study offers the first comparative look at the *same* developmental task across different online research methodologies, allowing for direct comparison and critical examination of each. We conducted three versions of a test of preschoolers' ability to generate and apply second-order inferences to predict novel outcomes. Experiment 1 is an in-person task conducted at public testing sites in the vicinity of the university. In Experiment 2, we conducted an online-moderated version of the same task, in which an experimenter presented a recording of the procedure during a live video call with families over Zoom. Finally, Experiment 3 is an online-unmoderated version of the task, in which the same videos were presented entirely asynchronously using the Lookit platform. Results suggest that online methodologies may introduce difficulties and age-related differences in young children's performance not observed in person. We consider these results in light of the previous online developmental replications, suggest possible interpretations, and offer initial recommendations to help future developmental scientists make informed choices about whether and how to conduct their research online.

Keywords: developmental research, internet, research methods, cognitive development, online research

INTRODUCTION

Much of modern behavioral psychology research is partially or entirely conducted online. The availability of survey creation software (Qualtrics, Gorilla, etc.) has enabled researchers to create digital experiments with relative ease. Online recruitment methods – including crowdsourcing platforms (Amazon Mechanical Turk, Prolific, etc.), social media, and messaging sites (Facebook,

Reddit, etc.), as well as online undergraduate participant pools maintained by universities – have allowed psychologists to expand the scope and scale of their research with considerably less effort and time than traditional methods. However, this sea change has occurred almost entirely within *adult* research. Despite being notoriously hampered by time-consuming and low-return recruitment methods, developmental psychology research has remained largely in-person. There are, of course, many legitimate reasons for this. In particular, developmental methods are daunting to digitize – participants are usually too young to read written instructions or text-based stimuli, studies are often highly interactive, and many involve manipulation of physical materials. In addition, the majority of systems and software developed for online research are designed to reach audiences 18 and older. In the absence of this infrastructure and faced with such unique challenges of translation and implementation, until recently, the majority of developmental psychology was conducted entirely off-line.

The recent and rapid move of developmental research onto online platforms can be attributed to two major developments. First, over the past 5–7 years, efforts to establish avenues of online research specifically designed for developmental science have begun to emerge. Researchers at MIT developed “Lookit,” the first large-scale crowdsourcing platform aimed at developmental populations and researchers (Scott et al., 2017; Scott and Schulz, 2017). Scientists can build studies within Lookit to record simple response and webcam data. These studies can then be made available to a large pool of families already registered on the Lookit website, and new families can also be invited to create accounts. At much the same time, Yale researchers launched TheChildLab.com (Sheskin and Keil, 2018), which aimed to more closely emulate traditional developmental methods by scheduling families for appointments with live researchers over video chat. During these sessions, experimenters present stimuli both verbally and visually using the video chat interface and can respond adaptively to participants and their parents in real-time.

Second, the widespread suspension of in-person activities due to COVID-19 created an urgent motivation to move developmental research online. In the last year, there has been a rapid acceleration in adoption and expansion of digital methodologies. As of early 2021, over 450 researchers from around 50 universities across seven countries were conducting research *via* Lookit,¹ and the majority of developmental research laboratories are now actively recruiting and testing participants *via* video chat platforms. Other unmoderated systems have also emerged, including discoveriesonline.org (operated by researchers at New York University, see Rhodes et al., 2020) and themusiclab.org (operated by Harvard University). Many of these researchers have also joined with others to form the ambitious project, CRADLE (Collaboration for Reproducible and Distributed Large-Scale Experiments; see Sheskin et al., 2020), which launched the joint website, ChildrenHelpingScience.com, as a centralized resource to house listings of online developmental research studies. As of June 2021, ChildrenHelpingScience.com includes over 800

studies from laboratories all over the world, roughly a third of which are intended for children under 6 years of age.

Empirical work on the validity of these platforms is still in its earliest stages, but findings published thus far are encouraging. Scott et al. (2017) conducted versions of three originally in-person experiments on Lookit, one each with infants (11–18 months), toddlers (24–36 months), and preschoolers (3- and 4-year olds). The latter task was a replication of Pasquini et al. (2007), which collected preschoolers’ verbal responses to investigate their sensitivity to the relative reliabilities of different informants. Although overall performance was lower on Lookit than in-person, the online study results followed the same general pattern across age groups and conditions as the original (Scott et al., 2017). In addition, Sheskin and Keil (2018) conducted several well-known developmental tasks using their video calling platform with children of different ages (5–6, 7–8, 9–10, and 11–12 years). The tasks spanned different domains, including memory (for number and size), social reasoning (fairness and false-belief), and physics reasoning (gravity). Children’s answers were largely consistent with expected in-person performance, except for the false-belief reasoning task, but even in this case, the pattern of results was significant (Sheskin and Keil, 2018). In addition, researchers at New York University have conducted successful conceptual replications of older children’s in-laboratory performance using online, unmoderated testing platforms (see Leshin et al., 2021 for a replication of the effects of generic language on essentialism in 4.5–8-year olds and; Nussenbaum et al., 2020 for a replication of the development of value-learning strategies in 8–25-year olds).

Notably, however, these studies have all sought to replicate in-person performance using a single online platform. There has not, as yet, been any research that compares the same developmental study *across* platforms. The options available for conducting developmental research online differ from one another in a variety of ways, and we do not yet know what effects, if any, these differences may have on children’s performance. Many of the practical trade-offs are readily apparent (more accessible, transparent, and efficient data collection, diversifying participant demographics, lower barriers to recruitment and participation, etc.; see Sheskin et al., 2020 for review). For example, although bypassing the need for real-time experimenters means that more initial effort is required to translate traditional methodologies to asynchronous platforms, this approach also reduces the considerable time, effort, and expertise usually required for collecting developmental data.

In contrast, the potential empirical consequences of these decisions are still largely unknown. Does the presence or absence of a real-time experimenter impact young children’s engagement with an online task? If so, how should this difference in engagement be weighed against the benefits of using prerecorded procedures in ensuring consistency across participants? Questions like these will be of vital importance for developmental science in the post-pandemic world. Thus, there is a growing need for data comparing these various platforms, which can enable developmental scientists to make informed choices about whether and how to conduct their research online.

¹<https://lookit.mit.edu/scientists/>

The current study offers the first comparative look at the different online research methodologies available to developmental science. We conducted three versions of the same task with preschoolers: Experiment 1 is an in-person task conducted at public testing sites in the vicinity of the university. In Experiment 2, we conducted an online-moderated version of the same task, in which an experimenter presented a recording of the procedure during a live video call with families over Zoom. Finally, Experiments 3a-c used an online-unmoderated version of the task, in which the same videos were presented entirely asynchronously using the Lookit platform. To our knowledge, this study is the first attempt to replicate the same developmental task across these three different methodologies, allowing for direct comparison and critical examination of each.

The task itself examines children's ability to generate and apply second-order inferences to predict novel outcomes. In contrast with first-order inferences, which focus on the concrete properties of objects and events, second-order inferences capture abstract relations among those objects and events. To illustrate this, imagine looking into the window displays of two storefronts. In the one on the left, you see shirts, pants, and sweaters, and on the right, shovels, clocks, and paintbrushes. The recognition of each individual item is a first-order inference – while the realization that all of the items within a particular window belong to the same higher-order category (“*clothes*,” on the left and “*tools*,” on the right) is second-order inference.

There is some evidence that the capacity for such higher-order inferences (e.g., that boxes contain objects that are the same shape) is present even in preverbal infants (Xu and Garcia, 2008; Dewar and Xu, 2010). However, this prior work has primarily looked at infants' *reactions* to events that are inconsistent with these second-order inferences (e.g., looking longer when a differently-shaped object is revealed). We do not know when learners begin to *utilize* these inferences to guide prediction and action. This capacity is a critical feature of second-order inferences in human reasoning. To return to our example, if you were asked which shop is more likely to sell umbrellas, you would likely be able to confidently recommend the shop on the right – despite never having observed this particular object in either window or knowing anything about the actual merchandise for sale inside.

Here, we ask whether children's inferences about unobserved populations are sensitive to the *variability* of observed samples and whether they can use this second-order information to predict which of two hidden populations is more likely to produce a novel outcome. To test this, children watched an experimenter randomly sample balls from two identical opaque containers. The *varied-sample* consisted of four differently colored balls, and the *uniform-sample* consisted of four identically colored balls. Children were then asked which of the two containers was more likely to contain a novel-colored ball inside. If children only consider these samples in terms of their first-order properties, then we would not expect them to show a preference for either container. Considering the samples' second-order properties, however, readily leads to an inference about the unseen populations involved. Thus, if children preferentially select the *varied-sample*

container, it would demonstrate that they have not only formed this second-order inference, but also are able to use it to guide their predictions and actions beyond the limits of their direct experience.

EXPERIMENT 1

In Experiment 1, we conduct an initial test of preschoolers' in-person performance on a *second-order inference* task. The experimental design and analysis plan were preregistered prior to beginning data collection.²

Participants

Forty children ($M = 40.12$ months, $SD = 5.12$ months, range = 25.35–47.8 months) were tested in Experiment 1 between November 2019 and March 2020. Participants were recruited and tested individually at local museums in a primarily urban area. While individual demographic information was not collected, demographics for recruitment locations suggest participants were predominately white (44.5%) and middle class (median household income of \$73,900).

A priori power analysis was performed to calculate the target sample size. Our effect size ($h = 0.72$) was based on results from Erb et al. (2013), which conducted a similar type of investigation (i.e., binomial analysis of a forced-choice inference question) with a similar age group. The minimum sample size needed to achieve a power of 0.8 at a significance level of 0.05 was 38, which we rounded to 40 to accommodate counterbalancing.

In addition, 13 children were tested but excluded and replaced due to experiment error ($n = 2$), sibling or caretaker interference ($n = 6$), or failure to respond to the test question ($n = 5$).

Stimuli

Two identical opaque containers (17" × 6" × 6") were constructed from black cardboard with a cardboard egg tray concealed inside. This tray allowed the experimenter to arrange the balls inside in a specific order and then identify and draw them without looking inside the box. A felt-covered opening at the top of each box allowed the experimenter to reach inside and draw the balls one at a time.

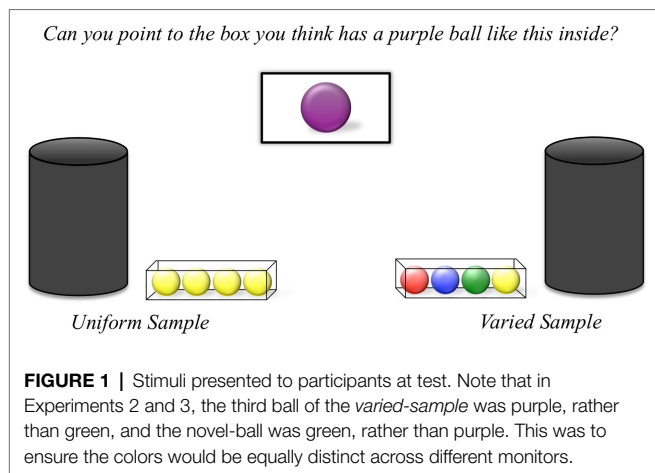
A total of 10 plastic golf balls of different colors were also used. These balls were placed inside of each of the two containers prior to the start of the task. One container held the *varied-sample*: one green, one red, one blue, and one yellow. The other container held the *uniform-sample* of four yellow balls. In addition, both containers held one *novel-ball*, which was purple.

The task also employed two 3" × 3" × 8" transparent plastic trays to hold the balls after they were drawn and a photograph of a single purple ball.

Procedure

Testing sessions began with the two opaque containers and clear trays on either side of the table (see **Figure 1**).

²<https://aspredicted.org/blind.php?x=rb4jn6>



The containers and trays were evenly spaced and equidistant from the participant. The experimenter told children they were going to play a game with the boxes, both of which had balls inside. She shook both containers so that the sound of the balls rattling inside was audible. The experimenter replaced the containers on the table and said, “I am going to show you some of the balls in each box,” and stepped to stand behind one of the two containers. The experimenter closed her eyes and turned her head away from the container while reaching in and pulling out a ball, apparently at random. She then directed her gaze toward the child while holding the ball out and said, “Look!,” before placing the ball into the clear plastic tray beside the container. This process of “sampling” was repeated three more times, for a total of four balls. Afterward, the experimenter repeated this process with the other container.

In this way, each participant observed a set of four balls drawn from each container. In the *uniform-sample*, all four were the same color (yellow), while in the *variable-sample*, the balls were all of different colors (one red, one blue, one green, and one yellow).³ The balls in the *variable-sample* were always drawn from the box in the same order. The order and side of presentation of the samples were counterbalanced across participants.

After drawing the second sample, the experimenter returned to the center of the table and addressed the child. Pointing at both the containers simultaneously, she said, “One of these two boxes has a purple ball, like this (holding a photograph of a purple ball), inside. Can you point to the box you think has the purple ball inside?” While asking this question, the experimenter looked straight ahead at the child to avoid biasing their response. If a child did not spontaneously indicate one of the two containers, the experimenter prompted by holding up the picture and repeating the question. Children who did not respond after two such prompts were excluded. After children indicated their choice, the experimenter reached into the selected container and drew a purple ball. Children were thanked for their participation and received a small gift.

³Samples were selected based on the procedure used by Sim and Xu (2013).

Results and Discussion

Children’s responses were recorded during the experimental session and videotaped. Response times were calculated as the time between the last word of the initial task question and when children initiated their response movement. The average response time was 8 s ($SD=7$ s, inter-rater reliability=90% of scores identical within ± 1 s), with only five children requiring repeated prompts to respond.

We recorded whether each child chose to search for the novel-colored ball in the *variable-sample* or the *uniform-sample* container. A significant majority of children (72.5%) chose the *varied-sample* container ($p=0.006$, two-tailed binomial). There was no significant effect of age on choice (Wald, $z=0.881$, $p>0.378$, *ns*). This suggests that young learners are not only able to form second-order inferences about the variability of unseen populations from the characteristics of observed samples, but can also apply this abstract property to guide subsequent predictions about novel events.

EXPERIMENT 2

Having demonstrated that preschoolers succeed on this task using a traditional, in-person procedure, Experiments 2 and 3 attempted to replicate this performance online. Experiment 2 conducted the task *via* an experimenter moderated video call with participants. Using a similar approach as Sheskin and Keil (2018), families who were interested in participating were directed to sign up for appointment slots (15 min each, primarily on weekend mornings) and guided through the study by an experimenter *via* video chat (Zoom).

Unlike previous work, however, the experimenter did not conduct the task herself. Instead, participants watched a video of another experimenter presenting the procedure used in Experiment 1. The “live experimenter” moderating the session controlled the playback of this video, pausing it at points when the child was asked to respond. This approach was chosen in order to maximize consistency of study delivery, which is one of the advantages of online research (e.g., Sheskin and Keil, 2018; Sheskin et al., 2020), without sacrificing the engagement and adaptability of presentation by a live experimenter. This also ensured consistency in study delivery across Experiments 2 and 3. While the ability to present online tasks in real time is a significant and potentially advantageous difference between moderated and unmoderated platforms, the goals of our investigation were best served by controlling this potential source of variation.

See Aspredicted.org for the preregistration of the experimental design and analysis plan for online replications.⁴ The video stimuli used in Experiments 2 and 3 can also be found at https://osf.io/5x8ku/?view_only=269c5468936d4811a55f237041f9ff96.

Participants

Online participants ($N=43$, $M=41.74$ months, $SD=3.47$ months, range=36.2–47.9 months) were tested in Experiment 2 between

⁴<https://aspredicted.org/blind.php?x=t85n33>

June and November of 2020. Children were recruited *via* email from a database of families maintained by the university's developmental laboratories. The majority of these were families who had previously been tested and/or indicated interest in future participation at an in-person testing site. Thus, participants in Experiment 2 were from roughly the same population as those in Experiment 1. In exchange for participating, families were offered a \$5.00 Amazon gift card.

An additional six children were tested, but excluded due to issues within the testing session: caretaker interference ($n=1$), failure to respond ($n=1$), or because technical issues or errors (unstable internet connection, etc.) interrupted the session ($n=4$).

Stimuli

Testing sessions were conducted *via* the Zoom video calling platform. Three prerecorded videos (*introduction*, *test*, and *conclusion*; described below) were presented to participants using Slides.com. This meant that participants accessed the videos directly *via* their own Internet connection, leading to fewer issues of lag than screen-sharing, while still allowing the experimenter to control video playback.

The only difference in the physical stimuli between Experiments 1 and 2 was a small change in the color of the balls. In order to ensure the colors would be distinguishable across different computer monitors, two colors were switched. The purple ball was used instead of green in the *varied-sample* and the *novel-ball* and corresponding picture card were green.

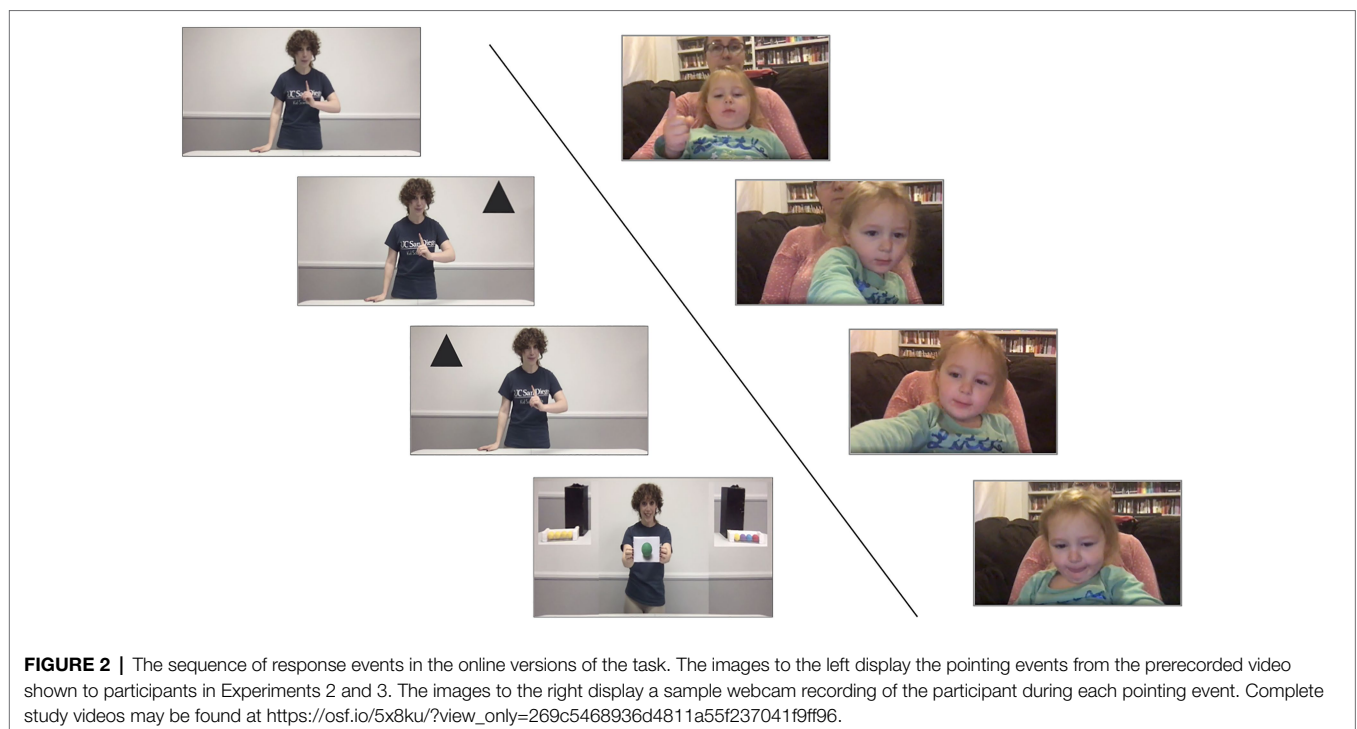
Procedure

Testing sessions began with the participating family joining the experimenter in a video call. This "live experimenter" would

introduce themselves to the parent and child and then give parents an overview of the session. Families were then sent a link to the Slides.com presentation *via* the video call chat function and parents were instructed to open it and full-screen the site window.

The live experimenter would then draw the child's attention to the screen and being playing the *introduction video*. This began with the "recorded experimenter" greeting the child and saying, "Before we start the game, let us practice using our pointing finger," while holding her hand out in front of her with index finger extended. A black triangle would then appear in either the top right or the top left corner of the screen (added in video-editing software post recording). The recorded experimenter asked children to use their pointing finger to "touch" the black triangle (pilot testing suggested that the best way to ensure a visually distinct "right/left" point was to instruct children to get close enough to touch the upper corner of their screens). The live experimenter would pause the video playback until the child had pointed and would repeat the instructions if needed. When the video continued, the recorded experimenter said, "Good job! Let us practice one more time," and the prompt was repeated with the triangle on the opposite side of the screen (left-right order counterbalanced across participants). This gave children a chance to practice the mode of response for the task and provided a visual calibration of what a choice for the left- or right-side container would look like (see **Figure 2**).

Next, the live experimenter advanced the presentation to the *test video*, in which the recorded experimenter performed the identical procedure from Experiment 1. At test, the recorded experimenter asked children to "touch" the box they thought contained the green ball. To ensure visibility of children's



responses, the images of the two boxes transitioned to the upper corners of the screen (see **Figure 2**). The live experimenter would pause the video until children responded. If children failed to respond spontaneously, the experimenter provided the same prompts as those used in Experiment 1. After providing a response, all children viewed a *conclusion video* in which the novel-colored ball was revealed from one container. The live experimenter then instructed parents to return to the video call window to conclude the session.

Results and Discussion

There was no significant difference in age between the participants tested in Experiments 1 and 2, $t(81) = -1.7$, $p = 0.09$ (*ns*). Children were somewhat more reluctant to respond to the task question in the moderated online platform than in person. The average response time was 11 s ($SD = 18$ s, inter-rater reliability = 91% of scores identical within ± 1 s), with 13 children requiring repeated prompts prior to responding.

The results of Experiment 2 showed a similar, but weaker pattern of performance observed in person: only 27 of the 43 children tested *via* video chat selected the *variable-sample* container (62.79%). Although this proportion was not significantly different from children's performance in Experiment 1 ($p = 0.213$, two-tailed binomial), it was also not significantly different from chance ($p = 0.126$, two-tailed binomial).

Post hoc analysis was conducted to see whether this non-replication might be due to age-related differences in online performance. A logistic regression treating age as a continuous factor was not significant (Wald, $z = 1.271$, $p > 0.204$, *ns*). However, a median-split of the sample revealed that children below 3.5 years of age ($n = 21$, $M = 38.62$ months, $SD = 1.74$ months, range = 36.2–41.5 months) selected the *varied-sample* container only 52.38% of the time, which was significantly less often than children 3.5 years of age and older ($n = 22$, $M = 44.71$ months, $SD = 1.49$ months, range = 42.21–47.9 months), who selected this container 72.73% of the time, $p = 0.048$, two-tailed binomial).

Given that the age and general population demographics of participants were the same between Experiments 1 and 2, the difference in results appears to be due to poorer performance of the youngest children in Experiment 2. Indeed, these results are similar to those reported by Scott et al. (2017), in which 3- and 4-year-olds' performance on Lookit was weaker than their in-person behavior, but showed the same general pattern. However, our results also suggest a developmental difference in online performance. Considering that younger children have necessarily had less experience interacting with online environments, it is possible that conducting the study online had a greater impact on their performance than older children. It is also possible that the online platform added noise equally across the age range and that younger children's second-order inference is simply less robust.

EXPERIMENT 3

Comparing children's performance in Experiments 1 and 2 suggests that online tasks that require an active behavioral

response (i.e., pointing) may impact performance, particularly for the youngest children. In Experiment 3, we expand this comparison to include an asynchronous online platform by using MIT's Lookit. This platform represents a greater departure from the characteristics of traditional developmental testing than online studies conducted over video calls. Interested families create accounts on the Lookit site and are notified of studies for available for their children's age range. The studies are composed of prerecorded and preprogrammed elements and are available for immediate participation at any time. We conducted three Lookit experiments: Experiment 3a and 3b sought to replicate the initial results with participants of the same age as those tested in Experiments 1 and 2. Then, Experiment 3c compares these results to the performance of slightly older children (4-years-old) on the same task.

Experiment 3a

Participants

A total of 41 children ($M = 41.88$ months, $SD = 3.55$ months, range = 36.39–47.77 months) were tested *via* Lookit between November and December of 2020. Demographic information collected from parents at the time they created their accounts indicates that participants were predominately white (75%) and upper-middle class (median household income of \$110,000). Families were offered a \$5 Amazon gift card for their participation.

An additional 28 children were excluded or dropped. The majority were children failing to respond to the test question ($n = 9$) or responding in a way that was not interpretable (e.g., pointing to the center of the screen, $n = 4$). In fewer cases, children responded too late to be fully recorded ($n = 3$) or children left during the videos ($n = 2$). The remaining 10 exclusions were due to technical errors disrupting either the presentation of stimuli ($n = 3$) or webcam recording ($n = 7$).

Stimuli

The prerecorded videos from Experiment 2 were used in Experiment 3. These videos were embedded into the Lookit platform, which automatically displayed them in a counterbalanced order. Webcam footage was recorded during the playback of each video. Prior to the videos, written instructions with images were presented to parents to explain how to set up for recording (see *Procedure*) and what to expect within the task.

Procedure

Testing sessions began when parents activated the study from the listing on the Lookit page. On the first screen of the study, written instructions outlined the task. Parents were then presented with a consent document and prompted to record a verbal consent video. This was followed by an opportunity to preview the actual study videos. If a parent chose to preview, they were directed to a new screen where they confirmed their child could not see the screen (webcam footage was also recorded during this preview to later confirm

the participant was not present). The preview video was a soundless, subtitled version of the task video, and presented with playback controls. All parents were then given instructions on how to set up for recording (single screen, centered webcam, etc.) and space (not backlit, faces clearly visible, minimizing distractions, etc.). Parents were instructed to have their child sit on their lap or beside them, but stressed that parents should not interact directly with their children during the game. Parents were provided with a preview of their webcam view to check that their child was visible and would be able to reach the screen, before advancing to the task itself.

A brief fixation video of a rotating ball played while webcam recording began. The task videos then played automatically. In order to ensure children had time to respond, the videos would automatically freeze for 20 s at each point where children were asked to respond (i.e., twice in the *introduction video* and once in the *test video*). Parents had the option to pause the task at any time, which would transition to a separate screen showing blank screen. After the *conclusion video*, parents read a debriefing script, which explained the purpose of the study and thanked them for their participation.

Results and Discussion

An analysis of variance showed no significant age differences between Experiment 3a and the previous two studies, $F(2,121)=2.315$, $p=0.103$ (*ns*). The average response time was 3 s ($SD=2$ s, inter-rater reliability=94% of scores identical within ± 1 s). However, as this only includes children who responded within the 20-s automated timeframe (see below), this response time cannot be readily compared to those in the previous two experiments.

As in Experiment 2, children's performance on Lookit showed a similar, but weaker trend as their in-person behavior. Overall, 26 out of 41 of children chose the *variable-sample* container at test (63.41%, $p=0.117$, two-tailed binomial). This was not significantly different from children's performance in either Experiment 1 ($p=0.22$, two-tailed binomial) or Experiment 2 ($p=1$, two-tailed binomial). However, unlike in Experiment 2, there were no significant age differences, either when age was treated as a continuous variable (logistic regression, Wald, $z=-1.159$, $p>0.246$) or when comparing the proportion of choices for children above and below 3.5-years-old ($p=0.269$, two-tailed binomial).

The rate at which children were excluded and replaced in this experiment (28 out of a total of 69) was markedly higher than either in-person (13 out of 53) or video chat (6 out of 43). Notably, the majority of exclusions were cases in which children did not respond within the automated timeframe provided for each question. This suggests that children who were faster to respond were more likely to be included in the final sample. It is therefore possible that our failure to replicate the in-person findings (Experiment 1) or age effects (Experiment 2) was due to this potential sampling bias. In an effort to correct this, a second Lookit experiment was designed to address this aspect of the initial design.

Experiment 3b

Although the length of response time provided in Experiment 3a (20 s) was substantially longer than children's average response times in person and on Zoom, it was insufficient for many children to respond on Lookit. In Experiment 3b, we therefore asked parents to advance the task manually after their child had responded. We also changed the implementation of webcam recordings to begin before the playback of the first video and end after the last one to capture all responses.

Participants

A total of 40 children ($M=41.1$ months, $SD=3.99$ months, range=36.07–47.80 months) were tested between February and April of 2021. Participants were predominately white (65.96%) and upper-middle class (median household income of \$130,000). Families were offered a \$3 Amazon gift card for their participation.

A total of 12 children were excluded. Very few children failed to respond at all ($n=2$) or provided uninterpretable responses ($n=4$). The remaining exclusions were all cases of technical errors disrupting the presentation of the stimuli ($n=6$).

Procedure

Aside from the change in manually advance the task, the procedure for Experiment 3b was identical to Experiment 3a. The video froze following the response prompts in the *introduction* and *test videos*, and a “next” button would appear. The video would remain paused until this button was clicked. In order to ensure that parents were aware of this aspect of the task, an additional instructions screen was added. This appeared just prior to the start of the *introduction video*.

Results and Discussion

An analysis of variance showed no significant age differences between this and the previous experiments, $F(3,160)=1.579$, $p=0.196$ (*ns*).

The changes in task implementation in Experiment 3b reduced exclusions. In Experiment 3a, 41% of children were excluded, and majority of those exclusions were due to their failure to respond to the task question in time. In Experiment 3b, the total rate of exclusion was reduced to 23%, with no children failing to respond. This rate of exclusion was much closer to that of the in-person study (24%). There was also greater variation in response time ($M=14$ s, $SD=28$ s, inter-rater reliability=89% of scores identical within ± 1 s), which is unsurprising given that responses were untimed, and there was no experimenter available to prompt children to respond (see section “General Discussion” for information on rates of parental “prompts” across all experiments).

Despite these improvements, however, we failed to replicate in-person performance. Overall, children responded at chance (57.5%, $p=0.43$, two-tailed binomial). This result was not different from performance in Experiment 2 ($p=0.515$) and Experiment 3a ($p=0.512$) and only marginally different from Experiment 1 ($p=0.049$, two-tailed binomial). As in Experiment 3a, there was no effect of age when treated as a continuous variable (logistic regression, Wald, $z=0.538$,

$p > 0.591$) or when comparing the proportion of *varied-sample* choices between younger and older children ($p = 0.521$, two-tailed binomial).

These results rule out the possibility that the lower performance observed in Experiment 3a was due to the time constraint on responses. However, it is unclear whether the lack of replication is due to an increased difficulty with the online task or whether the unfamiliar testing platform impeded children's second-order inferences. It is also possible that this digital, prerecorded context led children to treat the two samples as equivalent. In order to distinguish among these possibilities, we examine the online performance of slightly older children in Experiment 3c.

Experiment 3c

In an effort to identify what caused children's chance performance in the unmoderated online testing platform, we conducted another study on Lookit with an older sample of children. The results of Experiment 2 suggest that children's online performance on this second-order inference task may become more robust with age. If so, then older children's performance on an unmoderated online platform should be more likely to resemble in-person performance.

Participants

A total of 42 children ($M = 55.05$ months, $SD = 3.04$ months, range = 48.23–59.57 months) were tested during April of 2021. Demographic information indicated participating families were predominately white (61.36%) and middle class (median household income of \$100,000). Families were offered a \$3 Amazon gift card for their participation. Two additional children were excluded: one for inattention during the task videos and one for observing the study preview.

Method

The stimuli and task procedure were identical to Experiment 3b.

Results and Discussion

Of the 42 4-year olds tested in Experiment 3c, 32 selected the *varied-sample* container (76.19%). Performance was greater than in Experiment 3b ($p = 0.018$) and marginally greater than in Experiment 2 ($p = 0.08$), but not different from either Experiment 1 ($p = 0.73$) or Experiment 3a ($p = 0.108$, two-tailed binomial). See **Figure 3** for a comparison of children's performance across all experiments and platforms. As in Experiment 1, children choose the *varied-sample* significantly more often than chance ($p < 0.001$, two-tailed binomial).

Four-year-olds also responded much more readily to the task question, with an average response time of 4 s ($SD = 6$ s, inter-rater reliability = 97% of scores identical within ± 1 s), with no children failing to respond. Rates of parental involvement were also lower (see section "General Discussion").

These findings suggest that children are not making a genuinely different inference due to the online presentation of the study, but that younger children's ability to generate

and act on their inference may be less robust online than in-person.

GENERAL DISCUSSION

While still in their early stages, online platforms and protocols are poised to become a normalized and valuable part of developmental science. The COVID-19 pandemic forced researchers to meet the challenges of translating their studies into digital, distanced methodologies. Having overcome this initial hurdle, it is very likely that researchers will continue to utilize online methodologies after the return to in-person testing. The potential of online recruitment for accessing larger, more diverse, and lower-effort sources of developmental participants, as well as the ease of transparency, collaboration, and reproducibility of online protocols will continue to offer compelling opportunities for developmental science well into the post-pandemic world (Scott et al., 2017; Sheskin and Keil, 2018).

The current study offers an early, comparative look into how these possibilities might be realized across different online developmental research methods. We conducted the same second-order inference task with preschoolers in a traditional, in-person research setting (Experiment 1), *via* moderated video chat (Zoom; Experiment 2), and *via* an unmoderated, crowdsourcing site (Lookit; Experiments 3a–c). **Figure 3** compares children's performance across all experiments and platforms. In all versions of the task, the majority of children selected the *varied-sample* container, but this pattern of performance was weaker online. In both moderated and unmoderated online platforms, only the oldest children's (3.5–4-year old on Zoom and 4–5-year old on Lookit) choices were different from chance.

Considering our results in light of the previous findings suggests possible interpretations and recommendations for future developmental research. First, we were unable to fully replicate the children's successful in-person performance on a forced-choice second-order inference task in either moderated or unmoderated online platforms. This contrasts with previous research that has successfully replicated other in-person developmental results online. However, much of that work involved older children (e.g., Sheskin and Keil, 2018; Leshin et al., 2021) or implicit looking-time measures with infants (e.g., Scott et al., 2017). There is extensive evidence that children's success on looking-time measures precedes their ability to act in numerous domains (e.g., Zelazo et al., 1996; Hood et al., 2003; Kirkham et al., 2003). As such, it is perhaps unsurprising that preschoolers' performance on our task, which required an explicit response based on a second-order inference, was too fragile to translate online. We believe that these results should be treated as informative, rather than prohibitive, for future online research. They suggest that studies involving an explicit response from young children may be particularly challenging to conduct online, unless performance is expected to be particularly robust.

Notably, this research was conducted sequentially, rather than simultaneously, and this timing should be taken into

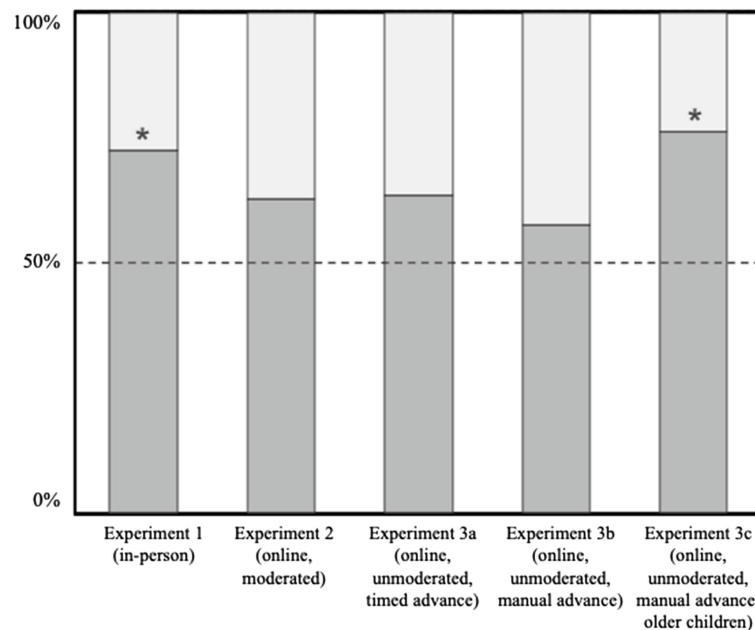


FIGURE 3 | The proportion of children's choices across in-person (Experiment 1), online-moderated (Experiment 2) and online-unmoderated (Experiments 3a–3c) testing platforms. Asterisks denote significance at $p < 0.05$.

account when interpreting the results. As noted in above, data collection for Experiment 1 was completed in early 2020, just prior to the stay-at-home orders due to COVID-19. All online testing was conducted over the course of the next 14 months: Zoom data collection for Experiment 2 began in the summer of 2020 and concluded in late fall, while the Lookit studies were conducted in late 2020 and early 2021. Thus, our data were collected during rather distinct periods of social and societal change. It is therefore difficult to speculate what impact these changes may have had on our results, especially since children's online performance does not seem to have improved with the dramatic increase in exposure to online platforms during this time.

This study also offers novel insights into the nature of online testing across different platforms, which may suggest important points of consideration for future research. For example, we found that parents were more inclined to interact and engage with their children during online testing sessions. Note that preregistered exclusion criteria prevented any potential impact of parental *interference* (e.g., a parent pointing at the stimuli). Non-interference interactions included neutral prompts and encouragements to respond (e.g., “Which one do you think it is?” “Can you point now?”). While these interactions were much more common during Zoom and Lookit testing than in-person (Experiment 1, $n = 3$), there was not much difference between the synchronous (Experiment 2, $n = 17$) and asynchronous untimed sessions (Experiment 3b, $n = 20$) for 3-year-old children. The rate of parental interactions was lower in Experiment 3a ($n = 9$), likely due to the limited response window, and in Experiment 3c ($n = 10$), which was conducted with 4-year-olds. This increased parental involvement

has potential to be beneficial, as it may help to reduce attrition during asynchronous testing. However, researchers should provide instructions to parents to control this interaction, and treat this aspect of the experiment as part of the study design.

Similarly, future researchers should make careful efforts to capitalize on the potential for online testing to broaden and diversify developmental participant pools. The current study did not attempt to control the demographics of Lookit participants and Experiments 3a–3c ultimately included *more* affluent, *less* diverse samples than those in Experiments 1 or 2. This will not only help to ensure the quality and comparability of online developmental data, but also to take advantage of the recruitment opportunities these platforms afford.

Finally, the current study highlights the potential use of online platforms for facilitating nuanced methodological and developmental comparisons. The time, effort, and cost of collecting developmental data often prohibit including additional comparison and control groups – even when this is the recommended approach. In the current study, we were able to conduct an identical version of a previous experiment with older children in order to clarify the developmental trajectory of children's performance on our task, with ease. While every effort was made to ensure the consistency of the procedure in Experiment 1, it was ultimately easier to achieve this consistency in Experiments 2 and 3. However, given the increased parental interaction, there was also some variability in the online procedures. Additionally, the period of data collection for the unmoderated online experiments (1–2 months) was less than half that of Experiments 1 and 2 (~5 months).

This study, along with the others in this special issue, represents some of the very first steps in better understanding the process, pitfalls, and potential of taking developmental research online. We hope that our results will serve to encourage and empower the field of developmental science to make the best possible use of these new methods going forward.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/5x8ku/?view_only=269c5468936d4811a55f237041f9ff96.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the University of California, San Diego, Human Research Protections Program. Written or video-recorded informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s) for the publication of any identifiable images or data included in this article.

REFERENCES

- Dewar, K. M., and Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: evidence from 9-month-old infants. *Psychol. Sci.* 21, 1871–1877. doi: 10.1177/0956797610388810
- Erb, C. D., Buchanan, D. W., and Sobel, D. M. (2013). Children's developing understanding of the relation between variable causal efficacy and mechanistic complexity. *Cognition* 129, 494–500. doi: 10.1016/j.cognition.2013.08.002
- Hood, B., Cole-Davies, V., and Dias, M. (2003). Looking and search measures of object knowledge in preschool children. *Dev. Psychol.* 39, 61–70. doi: 10.1037/0012-1649.39.1.61
- Kirkham, N. Z., Cruess, L., and Diamond, A. (2003). Helping children apply their knowledge to their behavior on a dimension-switching task. *Dev. Sci.* 6, 449–467. doi: 10.1111/1467-7687.00300
- Leshin, R. A., Leslie, S., and Rhodes, M. (2021). Does it matter how we speak about social kinds? A large, preregistered, online experimental study of how language shapes the development of essentialist beliefs. *Child Dev.* 92, e531–e547. doi: 10.1111/cdev.13527
- Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., and Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra Psychol.* 6:17213. doi: 10.1525/collabra.17213
- Pasquini, E. S., Corriveau, K. H., Koenig, M., and Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Dev. Psychol.* 43, 1216–1226. doi: 10.1037/0012-1649.43.5.1216
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (part 2): assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/OPMI_a_00001

AUTHOR CONTRIBUTIONS

EL and MG developed the concept, hypothesis, and design for the study. TT conducted the investigation and data curation, and drafted the manuscript. EL conducted the analysis and drafted and revised the manuscript. CW supervised the study, helped to develop the hypotheses, and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by funding from the National Science Foundation (CAREER grant #2047581), Hellman Foundation, Jacobs Foundation Fellowship, and the National Defense Science and Engineering Graduate Fellowship.

ACKNOWLEDGMENTS

The authors would like to thank Kim Scott and the Lookit Team for their assistance and support in conducting Experiment 3, the CRADLE team behind ChildrenHelpingScience.com, Trisha Katz, Phoebe Betts, Xiaoyang Chu, and Amberely Stein for their assistance in data collection, and the participating children and families, both in-person and online.

- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Sheskin, M., and Keil, F. (2018). TheChildLab.com a video chat platform for developmental research. *PsyArXiv [Preprint]*. doi: 10.31234/osf.io/rn7w5
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Sim, Z., and Xu, F. (2013). “Infants' early understanding of coincidences.” in *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 35.
- Xu, F., and Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proc. Natl. Acad. Sci. U. S. A.* 105, 5012–5015. doi: 10.1073/pnas.0704450105
- Zelazo, P. D., Frye, D., and Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cogn. Dev.* 11, 37–63. doi: 10.1016/S0885-2014(96)90027-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Lapidow, Tandon, Goddu and Walker. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Feasibility of Remote Performance Assessment Using the Free Research Executive Evaluation Test Battery in Adolescents

Isis Angelica Segura[†] and Sabine Pompéia^{*†}

Departamento de Psicobiologia, Universidade Federal de São Paulo, São Paulo, Brazil

OPEN ACCESS

Edited by:

Natasha Kirkham,
Birkbeck, University of London,
United Kingdom

Reviewed by:

Cynthia Whissell,
Laurentian University, Canada
Bhoomika Kar,
Allahabad University, India

*Correspondence:

Sabine Pompéia
spompéia@unifesp.br

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 10 June 2021

Accepted: 20 September 2021

Published: 14 October 2021

Citation:

Segura IA and Pompéia S (2021)
Feasibility of Remote Performance
Assessment Using the Free Research
Executive Evaluation Test Battery in
Adolescents.
Front. Psychol. 12:723063.
doi: 10.3389/fpsyg.2021.723063

Lockdowns and other preventive measures taken to curb the spread of diseases such as COVID-19 have restricted the use of face-to-face cognitive assessment. Remote testing may be an alternative, but it should first be shown to be comparable to in-person assessment before being used more widely, during and after the pandemic. Our aim was to evaluate the suitability of online, examiner-mediated administration of an open-access battery of executive function tests (the Free Research Executive Evaluation battery, or FREE) that can be adapted considering various characteristics of diverse populations and therefore used worldwide. A total of 96 9–15-year olds (42 girls) were tested, half of whom online through video calls mediated by an examiner. Their performance was compared to that of the other 48 individuals tested face-to-face, who were matched against the online-tested participants for age, pubertal status, sex, and parental schooling. The battery consists of two tests of the following executive domains: Updating (2-Back and Number Memory tests), Inhibition (Stroop Victoria and Stroop Happy-Sad), and Switching (Color Shape and Category Switch). Answers were vocal and self-paced, and the examiner recorded accuracy and time taken to complete in-person and online tasks. Only free software is needed for the assessment. Executive measures obtained from the tasks did not differ statistically between online and in-person tested participants and effects sizes of group effects were small, thus showing that the FREE test battery holds promise for online cognitive assessment, pending confirmation in different samples and further validation studies.

Keywords: adolescents, executive functions, COVID-19, online testing, updating, inhibition, shifting

INTRODUCTION

In the context of the COVID-19 pandemic's social distancing, researchers interested in cognition have looked at the feasibility of remote cognitive testing. Perhaps surprisingly, there is a substantial body of evidence from the last two decades showing that online cognitive assessment may be equivalent to lab-based, face-to-face testing (Krantz and Dalal, 2000; McGraw et al., 2000; Nosek et al., 2002; Temple et al., 2010; Soto et al., 2011; Germine et al., 2012; Cullum et al., 2014) in elderly (e.g., Geddes et al., 2020), adult (e.g., Kirkwood et al., 2000), and

pediatric populations (e.g., Hodge et al., 2019; Worhach et al., 2021). Although the validity and reliability of remote assessment has been questioned due to difficulties in controlling stimuli presentation and measuring response, both in terms of accuracy and reaction times (Germine et al., 2012), remote testing has unprecedented advantages that make it worth pursuing. As long as testees have internet access, these advantages include: (1) less travel time and expense, as well as lower implementation costs (Reips, 2000); (2) the possibility of reaching more diverse and less accessible samples, such as those from remote countries or areas, and/or patients with clinical conditions such as reduced mobility and/or higher vulnerability to diseases like COVID-19; and (3) testing people in familiar settings (their homes) instead of unknown locations, which has been shown to improve performance on some types of tasks (Sucksmith et al., 2013).

These advantages extend to testing BAME (Black, Asian and other non-white minority ethnic backgrounds) and non-WEIRD (Western Educated Industrialized Rich and Democratic) populations with internet access. These minorities or minoritized individuals have been under-represented in the cognitive literature in general, despite being more representative of humanity as a whole (see Henrich, 2010; Rad et al., 2018). Cultural, socioeconomic and ethnic differences affect not only the cognitive processes one might expect (e.g., social cognition and moral judgement) but also abilities such as visual perception, memory, categorization, attention, and executive functions (EF) (Henrich, 2010; Kelkar, 2013; Hackman and Gallop, 2015). EF is an umbrella-term for top-down cognitive functions that regulate an individual's behavior and emotions in order to achieve goals that are in peoples' minds (in their working memory) (Baggetta and Alexander, 2016; Friedman and Miyake, 2017). These behavioral self-regulation abilities have been found to be affected by many factors related to developing nations' poor, minority and minoritized groups (Moffitt et al., 2011), which have been hit harder by the pandemic, likely a longer-lasting threat for them due to a host of environmental factors (see Silva and Ribeiro, 2021).

In this scenario, remote EF cognitive testing must ensure health and safety of testees and examiners and also allow administration of cognitive tests that are adaptable to different cultural and socioeconomic contexts so as to more reliably capture the cognitive constructs under investigation (Fernández and Abe, 2018). Bearing this in mind, we investigated the adequacy of remote EF assessment mediated by examiners using a test battery designed to be adaptable to different contexts and populations (FREE: Free Research Executive Function Evaluation; Zanini, 2021).

The FREE test battery includes tasks that measure three types of executive functions that are interrelated, yet separable, based on a theoretical framework called the Unity and Diversity of Executive Functions (see Friedman and Miyake, 2017). These types of EF are inhibition of automatic responses, shifting between tasks, and updating information held in working memory. Importantly, these tasks were adapted for affordable testing by researchers using basic equipment. Task presentation and scoring are not automated. Testees themselves regulate the speed at which they can do tasks and respond vocally

while the examiner measures accuracy and time taken to complete blocks of trials instead of each individual trial. This is important because many studies have shown that limiting exposure and response times and requiring key presses for verbal answers can negatively influence measurements of EF in samples that include participants with different characteristics, such as various ages, and who are from diverse backgrounds (see Zanini, 2021). This test battery may therefore be administered remotely and be moderated by an online examiner, using screen sharing services that may be downloaded and used by examiners and testees free of charge without any special hardware, software downloads or plug-ins.

Here, the performance of adolescents tested online was compared with that of adolescents tested face-to-face in their own schools. Investigating EF is especially important during this phase of life because these cognitive skills develop during this period becoming differentiable in three distinct domains (see Lee et al., 2013), so factors that affect the environment and health at this age can potentially impact the development of EF performance, which influences a wide range of outcomes such as physical and mental well-being, academic and financial success, criminal and addictive behavior (see Moffitt et al., 2011). Hence, EF assessment of populations that include this vulnerable age is important. Additionally, it should be considered that many other factors reduce the possibility that people will be available for face-to-face testing. These include not only pandemics but also having debilitating illnesses that limit locomotion or the immune system, living in isolated locations, or others variables associated with poverty (e.g., shortage of means to pay for transport to and from research laboratories, or not having guardians who are available to accompany minors in person, which is often necessary). All of these conditions can also potentially impair EF, especially during sensitive phases of development like adolescence (Moffitt et al., 2011), which may go unnoticed if remote testing is not possible. Hunersen et al. (2021) call for the need to increase remote data collection strategies for testing adolescents, especially those from low-income settings using free tools, as proposed here.

The test battery used here was built for research purposes and group comparisons, not diagnostic evaluation, so norms are not available. For this reason, we matched participants tested online to others tested in person according to factors that are known to potentially affect brain and cognitive development (parental schooling, age, pubertal status, and sex: Foulkes and Blakemore, 2018). With the present study, we aim to establish whether applying the FREE battery is feasible under supervised online testing, to describe how the online testing was implemented and to compare the pattern of effects in both in-person and face-to-face conditions. Although completely remote, the supervised online presentation preserve some important aspects of in-person assessment that might be sources of bias if absent from online tests. These procedures included: (1) testee-examiner interaction throughout testing to prevent distraction from task objectives and misinterpreted instructions (Feenstra et al., 2017); (2) same format of stimuli presentation (PDF viewing of instructions and stimuli) that are seen (shared

screen when online) concomitantly by the testees and examiner; (3) same type of response to the tests (vocal, in which the examiner writes down the responses); and (4) self-paced format (clicking a mouse or tapping a keyboard to progress to the next stimulus during online testing and swiping the screen for in-person testing); the examiner used a stopwatch to time how long testees took to complete each task and wrote down the answers. Because the procedures were essentially the same except for being administered in person or online, we hypothesized that performance would be equivalent.

MATERIALS AND METHODS

Participants

Our convenience sample consisted of 96 native Portuguese-speaking, typically developing adolescents aged 9–16, of whom 48 were tested online. These individuals were matched to 48 adolescents who were evaluated in person at their schools (see matched pairing details below). Exclusion criteria were: (1) having been held back for a year or more at school; (2) being a student with special needs, which may be associated with clinical or cognitive limitations; and (3) taking daily medication to exclude any presence of chronic clinical disorders that could affect cognitive and/or developmental outcomes.

Procedures

This study was approved by the local Ethics Committee (CAAE # 56284216.7.0000.5505). Prior to the testing sessions, participants' guardians provided signed informed consent. Informed participant assent was always confirmed before test administration. Participants answered a socio-demographic questionnaire, self-evaluated their pubertal status filling in the 5-item Pubertal Development Scale [PDS, adapted from Carskadon et al., 1993 into Portuguese by Pompéia, 2019] and were administered the FREE executive function tasks in four pseudorandom orders to avoid the effects of fatigue. Participants from the supervised online group (hereafter called as "online group") were recruited through contact from their schools and social media after the authorities introduced social distancing and closed schools due to the COVID-19 pandemic. They were tested between 3 and 9 months after lockdown. All online participants were individually paired/matched to in-person tested individuals recruited at their schools and assessed the year before the pandemic broke out (some of whom were part of a prior study: Zanini, 2021). Matching was based on sex, age, pubertal status, and parent's average years of schooling as a proxy for cognitive stimulation, or socioeconomic status (SES; Sirin, 2005).

Both groups (online and in-person) were tested individually with supervision in a single session. Participants from the in-person group were tested at their schools using touchscreen tablets holding PDF files containing stimuli and instructions. Differently, participants tested online were at home and assessed remotely through an internet connection using their own hardware (computer or laptop, with web camera and

basic free software such as Adobe Acrobat, PowerPoint, and a free Zoom video communication application). These individuals were instructed by the experimenter to share their screens (step-by-step written and oral instructions were provided for those unfamiliar with Zoom). Next, participants were helped to download tests in PDF format from their own e-mails or their guardian's. These files were not available until testing started, so that they could not preview the tasks. For both groups, the examiner was present during the whole test session, supplying instructions, answering questions and ensuring participants were doing the tests as expected (e.g., not being interrupted by their cellphones and such like). All participants were awarded a "science partner" certificate after taking part and those tested in-person were reimbursed for their travel expenses. The EF test battery took around 40 min to be completed, including instructions and rest breaks if the participants required them. Approximate time taken to complete each task was: 2 min for both the Inhibition tasks, 4 min to complete the Color Shape task and 6 min to complete Category Switch (Shifting tasks), around 5 min to complete 2-Back task and 8 min to complete the Number Memory task (Updating tasks). Other tasks were administered to the same samples and their results will be reported elsewhere.

Cognitive Measures

The FREE battery contains six tests adapted for use in diverse samples in terms of SES and cultural context. The theoretical basis for the battery, the rationale for choice of tasks, description of tasks, answer sheets and scoring method are detailed in Zanini (2021). A brief description of tasks and scoring procedure for each domain can be found in **Table 1** and **Figure 1**. Following prior studies (see Zanini, 2021), the Inhibition and Switching tasks included blocks used to control for vocal/psychomotor speed and a corresponding block with the same requirements plus executive demands, while the Updating tasks contained no such control, as is the norm in this field.

To carry out the executive function tasks, testees read the instructions or had them read to them if preferred. The examiner clarified any questions that came up. Answers were vocal and the tasks were self-paced (swiping to pass from instructions to stimuli and between stimuli for the in-person group and clicking a mouse or tapping a forward keyboard arrow in the online group). The examiner wrote down the vocal answers and timed testees' task completions using a handheld stopwatch, akin to classic tests used to assess intelligence, for instance, following the long tradition of paper-and-pencil testing (for a further discussion on the advantages of this, see Kessels, 2019; Zanini, 2021). Sessions were recorded with participants' and their guardians' consent and erased once adequate scoring was ensured.

For each task or task block (depending on the test), the Rate Correct Score (RCS; see Vandierendonck, 2017) was calculated by dividing the number of correct answers by the time (in seconds) taken to complete each block. This metric controls for speed-accuracy trade-offs, that is, the

TABLE 1 | Description of the self-paced executive function tasks per domain and their corresponding scores (based on Zanini, 2021).

| Domain (Task) | Paradigm | Scores |
|--------------------|---|---|
| Inhibition | | |
| (Stroop Victoria) | Contains two blocks, each of which consists of 24 stimuli (color patches or words) displayed on a single screen. Participants name the ink color of patches (block 1) and words that are color names written in incongruous ink colors (block 2). Block 1 is the control block, measuring speed to name colors. Block 2 involves inhibition (naming ink colors of words instead of reading the color names) | Cost of inhibition: RCS in block 2 minus RCS in block 1 |
| (Happy-Sad Stroop) | Contains two scored blocks, each of which consists of 20 facial emotions that are displayed on a single screen. In block 1, participants name the emotions (happy or sad). In block 2, they inhibit naming the emotion and must name the opposite one (happy as sad or <i>vice-versa</i>) | Cost of inhibition: RCS in block 2 minus RCS in block 1 |
| Switching | | |
| (Color-Shape) | Contains three blocks in which single-colored geometric pictures are presented on each screen. As participants pass from screen to screen, pictures must be classified by shape (squares/circles) (block 1: 20 trials or screens), by color (black/gray) (block 2: 20 trials) or alternating (switching) classifications (block 3: 40 trials), according to cues presented above the pictures | Shifting costs: RCS in block 3 minus the sum of RCS in blocks 1 and 2 |
| (Category Switch) | Contains three blocks in which single pictures are presented on each screen. As participant pass from screen to screen, each pictures must be classified as living or non-living (block 1: 20 trials or screens), big or small (block 2: 20 trial) or alternating (switching) classifications without cues (living/non-living, then big/small and so forth) (block 3: 40 trials) | Shifting costs: RCS in block 3 minus the sum of RCS in blocks 1 and 2 |
| Updating | | |
| (2-Back) | Each screen contains 10 square outlines in fixed locations, one of which is filled in with black ink. As participants pass from screen to screen, they answer if the black square location they see is in the same or a different location as the black square two screens back. The total number of updating opportunities is 66 | Total RCS (for accuracy, in this case only: hits minus false alarms) (no control block) |
| (Number Memory) | Each screen contains a single digit number (1 to 9). As participants pass from screen to screen, they report the last three digits (trios) seen, in the same order as they were presented, having to continuously update information held in working memory, discarding the first digit of the trio and adding the new digit that appears next. The total number of updating opportunities is 24 | Total RCS (no control block) |

RCS = Rate Correct Score obtained by accuracy (vocal responses) divided by time (s) to complete blocks/task timed by the experimenter. See **Figure 1** for a visual illustration of the task.

between-participant variability in deciding to do tasks slowly, which can increase accuracy, or quickly, with higher error rates.

Statistical Analyses

Descriptive and inferential analyses were conducted using IBM SPSS Statistics version 21 software. The scores entered as dependent variables in the statistical analyses of the Inhibition and Shifting tasks were the executive costs (RCS from the block with executive requirements minus the RCS from control blocks). These dependent variables for each task were used in separate univariate General Linear Models (GLMs) with the factor group (online vs. in-person). For the Updating¹ measures, total RCS was used as the dependent variables in similar GLMs because they do not include baseline/control blocks. To correct for speed (of vocal responses and/or passing from stimulus to stimulus), our analyses included another continuous predictor, the mean RCS of the control blocks of the Inhibition and Shifting tasks, which have no executive requirements and basically involve answering aloud about an attribute of the stimuli (retrieval of phonological representations from long-term

memory) and speed of progressing through all stimuli, which is also done in the Updating task (see Zanini, 2021).

Because there is no prior data on online administration of the FREE test battery in the literature, calculating sample sizes was not strictly possible. We therefore focused on determining effect sizes (unstandardized and standardized) of the effect of group, which are useful to indicate the magnitude of the reported effects in metrics that are comparable across the tests. Standardized effect sizes are also useful to communicate the practical significance of the results, for meta-analysis and, importantly, can be used in future studies to determine sample size if these investigations intend to compare participants tested online and in-person (see Lakens, 2013). Effect sizes were presented as partial eta squared provided by SPSS and Hedges *g* [using Excel spreadsheet],² with corresponding confidence intervals. Rules of thumb describe medium effect sizes as those, respectively, between 0.059–0.138 and 0.5–0.8; values equal to or higher than 0.138 and 0.8 are considered large effect sizes (see Richardson, 2011; Lakens, 2013). Because participants for the different groups were matched by sex, age, pubertal stage, and SES, these variables were not supposed to be different between groups, so we did not include them as covariates in the analyses. Data were inspected for outliers (values over three SD of the mean).

¹For the 2-Back task, because participants can adopt a strategy of guessing (without actually updating the content in their working memory) by responding that all spatial configurations are different from the one presented two trials back (only 36% of trials are the same), accuracy in this case was calculated as hits minus false alarms (Jaeggi et al., 2010). Guessing in the other tasks can be picked up by the examiner but this never occurred in our experience.

²<https://www.cem.org/effect-size-calculator>

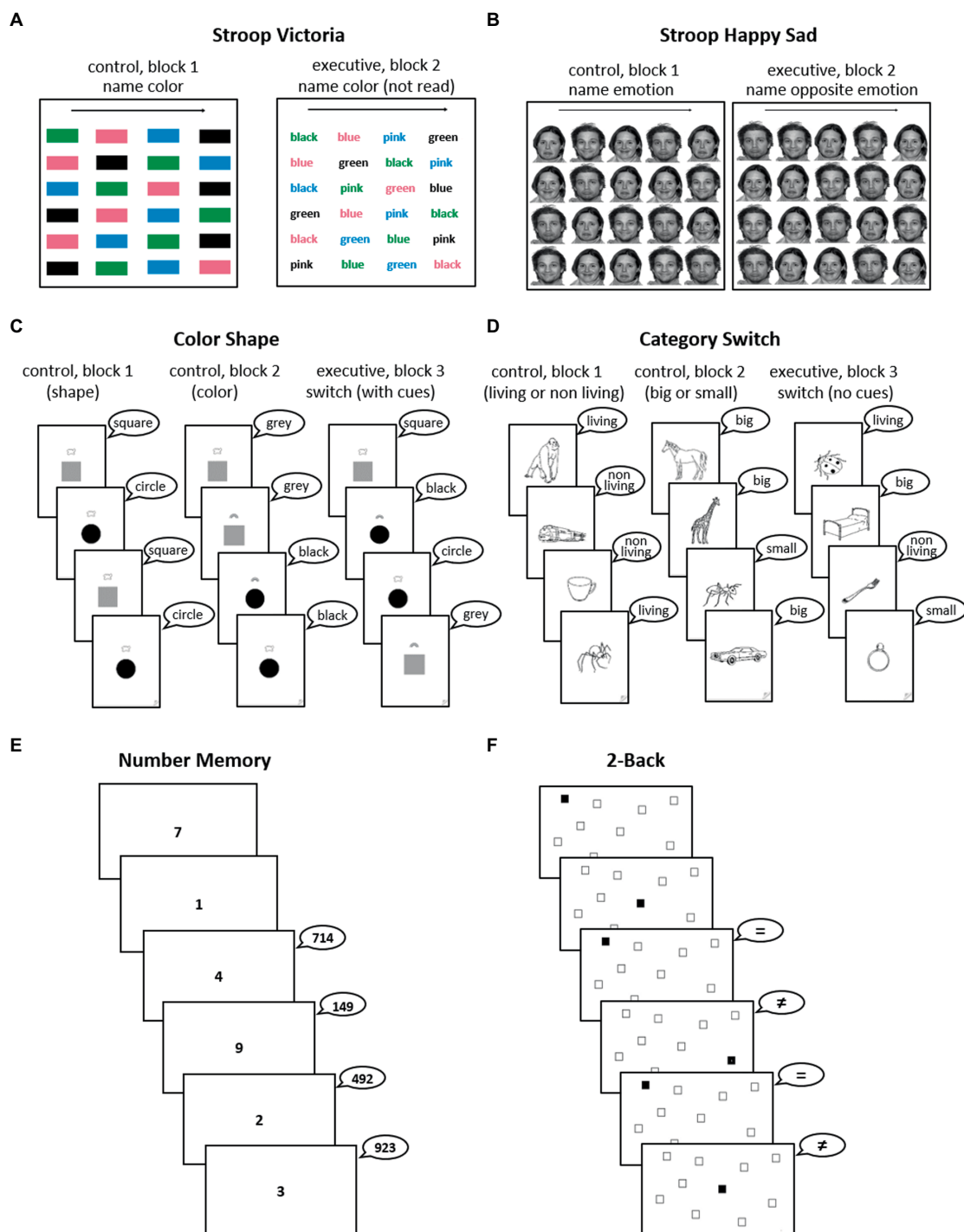


FIGURE 1 | Overview of the two tasks of each of the three executive domains: inhibition tasks (A,B), switching tasks (C,D) and updating tasks (E,F). In the Inhibition and Switching tasks the first blocks are the control blocks (naming characteristics of stimuli, with no executive requirements) and the last block requires executive abilities in addition to those involved in the control blocks. For details, see **Table 1** and Zanini (2021). All illustrated answers in speech bubbles are correct.

RESULTS

After matching participants tested online and in person (see **Table 1**), the proportion of participants of each sex, their age,

and pubertal status and their parents' average years of schooling were not different across groups. However, for one pair of matched participants there was a difference of 7 years in the mean of parental schooling. We therefore sought to further

control possible differences in cognitive stimulation between groups. As the in-person sample was mostly from public schools, except one, while all but one member of the online group were from private schools, the type of school (public vs. private) was included as a covariant in the analyses. This was done because Brazilian private-school students often outperform those from public schools (which tend to have lower quality education) on executive function tasks (e.g., Guerra et al., 2021).

The following outliers were found per variable: one in the Happy-Sad Stroop and one in the 2-Back for the in-person group and one in the Number Memory task for the online group. Both exclusion of these values or replacement for the value of the mean plus three SD retrieved similar results, so we report the results including data of these outliers.

Regarding type of test administration, we found no significant group effects (online vs. in person) in any of the executive tasks except for a marginal effect (small effect size) in the 2-Back test [$F(1,92)=3.899$, $p=0.051$, $p_{\eta^2}=0.04$], with a non-significant tendency for lower scores in the in-person group. No effects of type of school in any of the tasks were observed (see Table 2 and Figure 2), but there was an effect of the variable used to control for vocal/psychomotor speed in the 2-Back task [$F(1,92)=23.49$, $p<0.001$, $p_{\eta^2}=0.20$] and in the Number Memory task [$F(1,92)=55.09$, $p<0.001$, $p_{\eta^2}=0.37$], as expected, because the metrics used in the analysis of these tests do not have a baseline condition, unlike the inhibition and shifting tasks. The databank is available at https://osf.io/h5akr/?view_only=ea08777d698c46b4ae8110b9f8df8057t.

DISCUSSION

This study found no evidence that online, examiner-moderated use of the FREE test battery differs from in-person testing

(effect were not significant and of small effect sizes). This suggests that remote testing this way may be a comparable alternative when face-to-face assessment is not possible as found for other cognitive test batteries (Krantz and Dalal, 2000; McGraw et al., 2000; Nosek et al., 2002; Temple et al., 2010; Soto et al., 2011; Germine et al., 2012; Cullum et al., 2014). Hence, performing self-paced tasks and responding vocally, either personally or online, and using different hardware under these conditions did not affect results. This makes sense considering that the executive function variables were controlled for speed of vocal responses and passing from one stimulus to the next, irrespective of the conditions and equipment used by the participants: smaller screens and swiping to progress to the next stimuli during face-to-face assessment, or larger screens (laptops or personal computers) and mouse or key presses, remotely.

This absence of significant difference between in-person and remote testing may seem surprising considering that there can be minor delays and variation in precision when transmitting images, sounds and registering motor responses over the internet, although these seem to vary little across browsers, platforms, and operating systems (Anwyl-Irvine et al., 2020). The similar performances in the testing conditions used here may be explained by the FREE executive tests' design and scoring system, which is similar to classic paper and pencil tests, that have been used for decades and have considerable advantages over automated ones (Kessels, 2019). Time taken for each vocal answer is not the focus of interest in this test battery. Instead, it takes RCS into account, which is response accuracy divided by self-paced time taken for each task throughout a series of trials. Thus, total time to complete a task is much longer than reaction time per trial, which varies by milliseconds when computed in automatized tasks and could be affected by online transmission lags. Additionally, small variations in reaction

TABLE 2 | Descriptive statistics of demographics and Rate Correct Scores (RCS: accuracy divided by total time in s) of executive functions performance.

| Variables | In person (n=48; 21 girls) | Online (n=48; 21 girls) | Hedges g (95% CI) | Group effects | | | Type of School effects | | |
|---|-------------------------------|----------------------------|----------------------|---------------|-------|-----------------------------|------------------------|------|-----------------------------|
| | Mean (±SD) | Mean (±SD) | | F | p | p _η ² | F | p | p _η ² |
| Demographics | | | | | | | | | |
| Age (years) | 12.29 (1.97) | 12.17 (1.96) | | | | | | | |
| PDS (score) | 2.37 (0.71) | 2.27 (0.71) | | | | | | | |
| Guardian's schooling (\bar{x} yrs.) | 14.78 (2.27) | 14.86 (2.58) | | | | | | | |
| Executive functions | | | | | | | | | |
| Inhibition | | | | | | | | | |
| Stroop Victoria (inhibition cost) | −0.66 (0.28) | −0.54 (0.27) | −0.43 (−0.84/−0.03) | 2.239 | 0.14 | 0.02 | 0.016 | 0.9 | <0.01 |
| Stroop Happy-sad (inhibition cost) | −0.44 (0.29) | −0.45 (0.27) | 0.04 (−0.36/0.44) | 0.878 | 0.88 | 0.01 | 0.728 | 0.40 | 0.01 |
| Switching | | | | | | | | | |
| Color-Shape (switching cost) | −0.45 (0.18) | −0.47 (0.17) | 0.11 (−0.29/0.51) | 0.028 | 0.87 | 0.001 | 0.038 | 0.85 | <0.01 |
| Category Switch (switching cost) | −0.33 (0.15) | −0.38 (0.14) | 0.34 (−0.06/0.74) | 1.525 | 0.22 | 0.02 | 0.009 | 0.93 | <0.01 |
| Updating (corrected for baseline speed)* | | | | | | | | | |
| 2-Back | 0.27 (0.17) | 0.37 (0.15) | −0.62 (−1.03/−0.21) | 3.899 | 0.051 | 0.04 | 0.147 | 0.70 | <0.01 |
| Number Memory | 0.18 (0.05) | 0.17 (0.05) | 0.20 (−0.20/0.60) | 0.884 | 0.35 | 0.01 | 0.000 | 0.99 | <0.01 |

PDS=Pubertal Development Scale; *baseline speed effect for the 2-Back $F(1,92)=23.49$, $p<0.001$, $p_{\eta^2}=0.20$; Number Memory $F(1,92)=55.09$, $p<0.001$, $p_{\eta^2}=0.37$.

times of examiners, who mark the beginning and end of each block of trials, are comparable in the control and executive conditions/blocks, online and in person, and become irrelevant when compared to total time taken to complete each block/task. Although this presents some extra workload for examiners, it allows testing under a wider variety of conditions (Kessels, 2019). These tasks can therefore be used by examiners who do not have access to special hardware and software that automatically time responses for each trial and in private practices or poorly equipped laboratories, given that most of the world's neuropsychologists cannot afford these gadgets and applications. The present study shows that this advantage extends to online testing, as long as it is mediated by an examiner. In effect, inconsistent results have been found when comparing performance in tasks that are self-administered (e.g., in hospital vs. online settings: Feenstra et al., 2017), possibly because cognitive testing often needs supervision. Mediated testing allows examiners to make sure that testees understand tasks, pay attention when doing them and do not engage in the use of strategies that may distort performance.

We did find, however, a marginal group effect in the 2-Back task, which should be addressed. We envisage two possible explanations for this: (1) it could have been a spurious effect; and (2) although we controlled for type of school, it is possible that this statistical adjustment did not correct for a putative advantage of the online group, which had access to better schooling, having all been from private schools except for one individual. Indeed, Guerra et al. (2021) have recently shown that children from private schools in Brazil had higher spans for spatial locations (however, of a small effect size), but that spatial updating task performance, adjusting for span, did not differ between public and privately schooled individuals. Consequently, possible slightly higher spans due to higher cognitive stimulation could have contributed to the 2-Back results in the online group, which was not assessed here because of the nature of our 2-Back task, which differs from the n-back test of Guerra et al. (2021). In line with this idea, our findings show that the marginal non-significant effect in the 2-Back task does not seem to be specific to updating-related executive functioning, because group effects in the other task of this domain (Number Memory) were nowhere near significant, nor close to a medium effect size. It is also of note that, despite the lack of norms, the marginally significant difference found between scores from the in-person and online groups in the 2-Back task seem to have been due to a lower performance of the in-person group ($\text{mean} \pm \text{SD } 0.27 \pm 0.17$) because the online group (0.37 ± 0.15) presented results in this task that are very similar (0.35 ± 0.13) to those found by Zanini (2021), in which all participants were tested in person. Indeed, overall, performance of the online sample was very similar to that obtained in a comparable population using the same battery in person: means obtained here were within $\text{mean} \pm 0.5$ SD of Zanini (2021), which correspond to low effect size differences (Lakens, 2013).

The FREE test battery was designed with two tasks of three executive domains, so that consistency of effects between each pair of tests of each type can be used to ascertain the influence

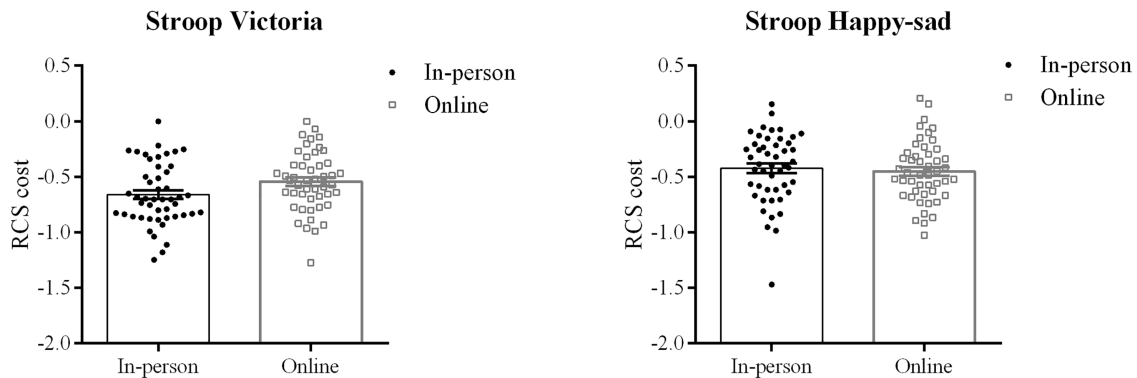
of various factors such as mode of testing (online vs. in-person), sex, SES, and so forth. Having two tasks of each domain also allows latent factors to be estimated, considering many multifactor configurations found for the Unity and Diversity Model of Executive Functions (see Karr, 2018). Unlike raw scores used here, latent factors capture the common variance in performance across different tasks, free of measurement errors (Brown et al., 2015). Therefore, specific cognitive requirements of a task that are not shared with the corresponding test of the same domain (such as spatial span in the 2-Back but not the Number Memory task) should not contribute to the latent factor. Our sample was too small to explore the latent nature and best model configuration of the executive functions' Unity and Diversity model, and to obtain evidence of invariance (Meade and Bauer, 2007) across mode of testing (online vs. in-person), so this must be undertaken in future studies. Another advantage of having a couple of tests per domain is that researchers who intend to use only one task of each EF type can pick the one which they deem more adequate for their purposes, although the ideal is to obtain latent scores. This, however, is only possible with large samples.

Considering the present scenario, evidence is emerging that COVID-19 and similar infections may lead to cognitive problems (cognitive COVID) beyond the acute stage (Ritchie and Chan, 2021), so repeatedly assessing infected patients may become essential to understand possible long-term cognitive impact. This includes children and adolescents, some of whom present long-term COVID effects (see Hertting, 2021) and will need to be followed up. As we have shown here, this may be done remotely with supervision, enabling a greater number of patients to be assessed if they have internet access, some familiarity with digital tools, a computer or device that runs basic software, a web camera, and a reasonably sized screen. All tests used may be easily adapted for diverse populations and are affordable (see Zanini, 2021) for poorly funded researchers.

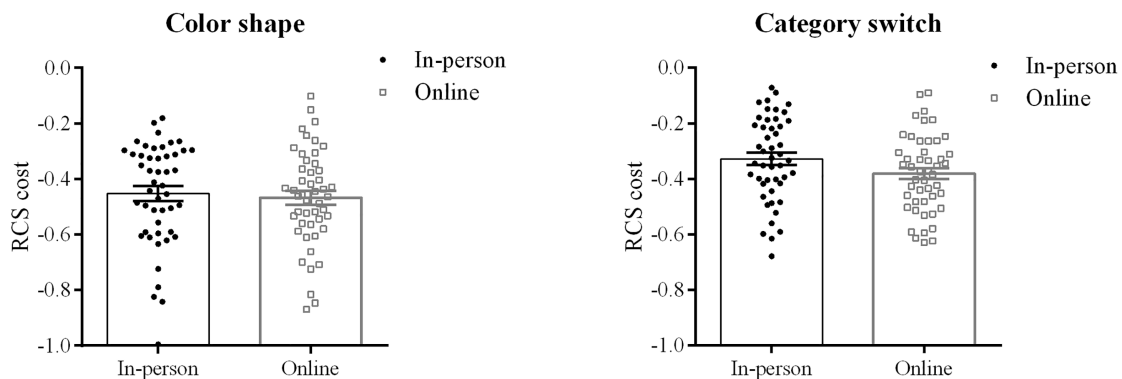
Additionally, a point to bear in mind is that online testing may pose technical issues such as unreliable connections or slow speeds. Testees may also not be comfortable with the technology involved or not have the hardware and an internet connection, which might be the case for those from low SES. Furthermore, examiners may fail to notice difficulties that are not clear through vocal communication when participants are unable or unwilling to turn on their cameras (common in youngsters: Castelli and Sarvary, 2021). Online testing also poses some ethical problems that must be minimized such as violating privacy. Nonetheless, despite these shortcomings, online testing probably reaches much larger numbers at low cost and will therefore probably become more prevalent in the post-pandemic period.

The main limitation of the present study is that participants in the in-person group were tested before the pandemic and those tested online were assessed during the pandemic, so this could have affected the results. Ideally, both groups should have been evaluated in parallel, but due to the pandemic this was not ethically acceptable. Nonetheless, the negative acute effects of the lockdown, which seem to be more severe (Creswell et al., 2021), were avoided, as we only tested participants from

Inhibition tasks



Switching tasks



Updating tasks

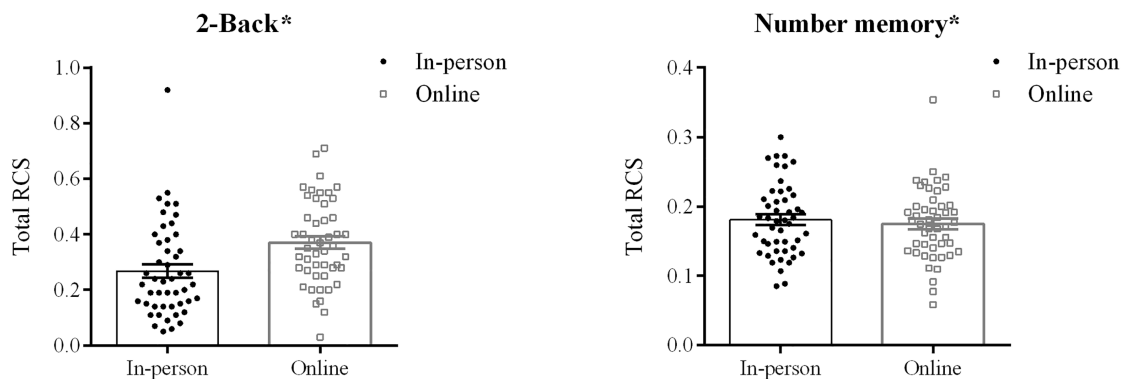


FIGURE 2 | Individual (dots) and mean (\pm SE) scores (histograms with error bars) per type of test administration [participants tested in person (black circles) and online (open grey squares)] in each of the executive function measures. *data shown without correction for speed.

3 months after the beginning of the social distancing measures. Additionally, potential participants who were reported by guardians as not being healthy were not tested in either group. If the COVID-19 crisis had affected participants in ways that

interfered with their EF, beyond those that were controlled for by matching participants (which are the known factors that influence these cognitive abilities: Foulkes and Blakemore, 2018), it would be expected that the online group perform

worse, which was not observed. Various factors that could have changed due to the pandemic, such as increases in mental health problems, sedentarism, body weight, alterations in sleep patterns, and so forth are not usually considered at all in prior studies that investigated the same EF model as used here (e.g., Friedman et al., 2011, who used the same sample of twins in many publications). Hence, it would be unclear how or whether they could have affected results. Another limitation of our study includes its sample size. Because our aim was to test feasibility of supervised remote online testing using the FREE battery, we did not have a sample large enough to allow us to run confirmatory factor analysis of the Unity and Diversity model of executive functions (see Friedman and Miyake, 2017; Karr, 2018) and to perform invariance testing and other types of validation techniques to ensure that web-based assessment was tapping the same constructs as those measured face-to-face (see Germine et al., 2012). Matching participants not only by parental schooling but also the type of school would have also been ideal. We attempted to control for the latter statistically, but it might have not been an effective control as almost all participants in the online group were from private schools, which more readily agreed to help volunteers participants during school closure. We had little success accessing families through public schools because their staff were much more severely overburdened due to the pandemic and extremely low governmental funding to aid the transition to online teaching. Finally, unlike other similar studies that investigated the adequacy of remote cognitive assessment by controlling for individual differences in performance using within-participant designs (e.g., Cullum et al., 2014; Feenstra et al., 2017; Backx et al., 2020), we did not test the same participant face-to-face and online because test-retest reliability of executive functions is known not to be high (Karlsen et al., 2020) because testees develop strategies that minimize executive functioning. On the upside, this experiment used a sample from a developing nation, which is still rare in the international literature (Rad et al., 2018), especially regarding the adequacy of remote cognitive assessment.

Overall, although our findings cannot be generalized to samples from other cultures and age groups, we have shown that online testing using the FREE test battery is a potentially viable means of remotely assessing EF. Because this test battery is open access, adaptable to populations with different characteristics and remote testing was done using only free software, our results provide initial evidence for a much-needed remote way to assess adolescents at low costs, including those who are more vulnerable to factors that negatively affect developmental (Hunersen et al., 2021; Moffitt et al., 2011), including BAME and non-WEIRD populations, who are under-represented in the cognitive literature (see Henrich, 2010;

Rad et al., 2018). We conclude that online testing the way it was administered here is a feasible way of collecting data on EF, making this a potential alternative when face-to-face testing is not possible. Until more controlled experiments are conducted, it is advisable to either test all participants online or in person and not mix these conditions, which was not assessed here.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found at: https://osf.io/h5akr/?view_only=ea08777d698c46b4ae8110b9f8df8057t.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Comitê de Ética em Pesquisa of the Universidade Federal de São Paulo. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SP and IS: designing the study, planning of the study, data collection, data analysis, and interpretation of the data. SP: public responsibility for the content of the article. All authors contributed to the article and approved the submitted version.

FUNDING

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP – Processes: due to fellowships to author IS 2019/19709–6 and SP 2016/14750–0), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (finance code 001), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – #301899/2019–3 due to fellowships to author SP), and Associação Fundo de Incentivo à Pesquisa (AFIP).

ACKNOWLEDGMENTS

We thank Rafaella Sales de Freitas, Luanna Inácio, Eveli Truksinas, Robson Kojima, Beatriz Zappellini, Daniel Utsumi, Mariana Libano, Bianca Rodrigues and Diogo Marques for help with data collection.

REFERENCES

- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., and Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behav. Res. Methods* 53, 1407–1425. doi: 10.3758/s13428-020-01501-5
- Backx, R., Skirrow, C., Dente, P., Barnett, J. H., and Cormack, F. K. (2020). Comparing web-based and lab-based cognitive assessment using the Cambridge neuropsychological test automated battery: a within-subjects counterbalanced study. *J. Med. Internet Res.* 22:e16792. doi: 10.2196/16792
- Baggetta, P., and Alexander, P. A. (2016). Conceptualization and operationalization of executive function. *Mind Brain Educ.* 10, 10–33. doi: 10.1111/mbe.12100

- Brown, G. T. L., Harris, L. R., O'Quin, C., and Lane, K. E. (2015). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: identifying and understanding non-invariance. *Int. J. Res. Method. Educ.* 40, 66–90. doi: 10.1080/1743727X.2015.1070823
- Carskadon, M. A., Vieira, C., and Acebo, C. (1993). Association between puberty and delayed phase preference. *Sleep* 16, 258–262. doi: 10.1093/sleep/16.3.258
- Castelli, F. R., and Sarvary, M. A. (2021). Why students do not turn on their video cameras during online classes and an equitable and inclusive plan to encourage them to do so. *Acad. Pract. Ecol. Evol.* 2020, 3565–3576. doi: 10.1002/ece3.7123
- Creswell, C., Shum, A., Pearcey, S., Skripkauskaitė, S., Patalay, P., and Waite, P. (2021). Young people's mental health during the COVID-19 pandemic. *Lang. Child Adolesc. Health* 5, 535–537. doi: 10.1016/S2352-4642(21)00177-2
- Cullum, C., Munro, L., Hynan, S., Grosch, M., Parikh, M., and Weiner, M. F. (2014). Teleneuropsychology: evidence for video teleconference-based neuropsychological assessment. *J. Int. Neuropsychol. Soc.* 20, 1028–1033. doi: 10.1017/S1355617714000873
- Feenstra, H. E. M., Murre, J. M. J., Vermeulen, I. E., Kieffer, J. M., and Schagen, S. B. (2017). Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam cognition scan. *J. Clin. Exp. Neuropsychol.* 40, 253–273. doi: 10.1080/13803395.2017.1339017
- Fernández, A. L., and Abe, J. (2018). Bias in cross-cultural neuropsychological testing: problems and possible solutions. *Culture. Brain* 6, 1–35. doi: 10.1007/s40167-017-0050-2
- Foulkes, L., and Blakemore, S. J. (2018). Studying individual differences in human adolescent brain development. *Nat. Neurosci.* 21, 315–323. doi: 10.1038/s41593-018-0078-4
- Friedman, N. P., and Miyake, A. (2017). Unity and Diversity of executive functions: individual differences as a window on cognitive structure. *Cortex* 6, 186–204. doi: 10.1016/j.cortex.2016.04.023
- Friedman, N. P., Miyake, A., Robinson, J. L., and Hewitt, J. K. (2011). Developmental trajectories in toddlers' self-restraint predict individual differences in executive functions 14 years later: a behavioral genetic analysis. *Dev. Psychol.* 47:1410. doi: 10.1037/a0023750
- Geddes, M. R., O'Connell, M. E., Fisk, J. D., Gauthier, S., Camicioli, R., and Ismail, Z. (2020). Remote cognitive and Behavioral assessment: report of the Alzheimer Society of Canada task force on dementia care best practices for COVID-19. *Alzheimer's. Dementia: Diagnosis. Assess. Dis. Monit.* 12, 1–11. doi: 10.1002/dad2.12111
- Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychon. Bull. Rev.* 19, 847–857. doi: 10.3758/s13423-012-0296-9
- Guerra, A., Hazin, I., Guerra, Y., Roulin, J. L., Le Gall, D., and Roy, A. (2021). Developmental profile of executive functioning in school-age children From Northeast Brazil. *Front. Psychol.* 11:596075. doi: 10.3389/fpsyg.2020.596075
- Hackman, D. A., Gallop, R., Evans, G. W., and Farah, M. J. (2015). Socioeconomic status and executive function: developmental trajectories and mediation. *Dev. Sci.* 5, 686–702. doi: 10.1111/desc.12246
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–135. doi: 10.1017/S0140525X0999152X
- Hertting, O. (2021). More research is needed on the long-term effects of COVID-19 on children and adolescents. *Acta Paediatrica. Int. J. Paediatrics.* 110, 744–745. doi: 10.1111/apa.15731
- Hodge, M. A., Sutherland, R., Jeng, K., Bale, G., Batta, P., Cambridge, A., et al. (2019). Agreement between telehealth and face-to-face assessment of intellectual ability in children with specific learning disorder. *J. Telemed. Telecare* 25, 431–437. doi: 10.1177/1357633X18776095
- Hunersen, K., Ramaiya, A., Yu, C., Green, J., Pinandari, A. W., and Blum, R. (2021). Considerations for remote data collection among adolescents During the COVID-19 pandemic. *J. Adolesc. Health* 68, 439–440. doi: 10.1016/j.jadohealth.2020.11.020
- Jaeggi, S. M., Studer-Luethi, B., Buschkuhl, M., Yi Fen, S., Jonides, J., and Perrig, W. J. (2010). The relationship between N-Back performance and matrix reasoning-implications for training and transfer. *Intelligence* 38, 625–635. doi: 10.1016/j.intell.2010.09.001
- Karlsen, R. H., Karr, J. E., Saksvik, S. B., Lundervold, A. J., Hjemdal, O., Olsen, A., et al. (2020). Examining 3-month test-retest reliability and reliable change using the Cambridge neuropsychological test automated battery. *Appl. Neuropsychol. Adult.* 21, 1–9. doi: 10.1080/23279095.2020.1722126
- Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., Karr, J. E., et al. (2018). The Unity and Diversity of executive functions: a systematic review and re-analysis of latent variable studies The Unity and Diversity of executive functions: a systematic review and re-analysis of latent variable studies. *Psychol. Bull.* 144:1147. doi: 10.1037/bul0000160
- Kelkar, A. S., Hough, M. S., and Fang, X. (2013). Do we think alike? A cross-cultural study of executive functioning. *Culture. Brain* 1, 118–137. doi: 10.1007/s40167-013-0010-4
- Kessels, R. P. (2019). Improving precision in neuropsychological assessment: bridging the gap between classic paper-and-pencil tests and paradigms from cognitive neuroscience. *Clin. Neuropsychol.* 33, 357–368. doi: 10.1080/13854046.2018.1518489
- Kirkwood, K. T., Peck, D. F., and Bennie, L. (2000). The consistency of neuropsychological assessments performed via telecommunication and face to face. *J. Telemed. Telecare* 6, 147–151. doi: 10.1258/1357633001935239
- Krantz, J. H., and Dalal, R. (2000). "Validity of web-based psychological research," in *Psychological Experiments on the Internet*. ed. M. H. Birnbaum (California: Academic Press), 35–60.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Front. Psychol.* 4:863. doi: 10.3389/fpsyg.2013.00863
- Lee, K., Bull, R., and Ho, R. M. (2013). Developmental changes in executive functioning. *Child Dev.* 84, 1933–1953. doi: 10.1111/cdev.12096
- McGraw, K. O., Tew, M. D., and Williams, J. E. (2000). The integrity of web-delivered experiments: can you trust the data? *Psychol. Sci.* 11, 502–506. doi: 10.1111/1467-9280.00296
- Meade, A. W., and Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Struct. Equ. Model. Multidiscip. J.* 14, 611–635. doi: 10.1080/10705510701575461
- Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., et al. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proc. Natl. Acad. Sci.* 108, 2693–2698. doi: 10.1073/pnas.1010076108
- Nosek, B. A., Mahzarin, R., and Greenwald, A. G. (2002). E-research: ethics, security, design, and control in psychological research on the internet. *J. Soc. Issues* 58, 161–176. doi: 10.1111/1540-4560.00254
- Pompéia, S., de Almeida Valverde Zanini, G., Inacio, L. M. C., da Silva, F. C., de Souza Vitale, M. S., Niskier, S. R., et al. (2019). Adapted version of the pubertal development scale for use in Brazil. *Rev. Saude Publica* 53:56. doi: 10.11606/s1518-8787.2019053000915
- Rad, M. S., Martingano, A. J., and Ginges, J. (2018). Toward a psychology of homo sapiens: making psychological science more representative of the human population. *Proc. Natl. Acad. Sci.* 115, 11401–11405. doi: 10.1073/pnas.1721165115
- Reips, U.-D. (2000). The web experiment: advantages, disadvantages, and solutions. *Psychol. Exp. Internet.* 1995, 89–117.
- Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educ. Res. Rev.* 6, 135–147. doi: 10.1016/j.edurev.2010.12.001
- Ritchie, K., and Chan, D. (2021). The emergence of cognitive COVID. *World Psychiatry* 20, 52–53. doi: 10.1002/wps.20837
- Silva, J., and Ribeiro, M. (2021). Short report social inequalities and the pandemic of COVID-19: the case of Rio de Janeiro. *J. Epidemiol. Community Health* 75, 1–5. doi: 10.1136/jech-2020-214724
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: a meta-analytic review of research. *Rev. Educ. Res.* 75, 417–453. doi: 10.3102/00346543075003417
- Soto, C. J., John, O. P., Gosling, S. D., and Potter, J. (2011). Age differences in personality traits From 10 to 65: big five domains and facets in a large cross-sectional sample. *J. Pers. Soc. Psychol.* 100, 330–348. doi: 10.1037/a0021717
- Sucksmith, E., Allison, C., Baron-Cohen, S., Chakrabarti, B., and Hoekstra, R. A. (2013). Empathy and emotion recognition in people with autism, first-degree relatives, and controls. *Neuropsychologia* 51, 98–105. doi: 10.1016/j.neuropsychologia.2012.11.013

- Temple, V., Drummond, C., Valiquette, S., and Jozsvai, E. (2010). A comparison of intellectual assessments over video conferencing and in-person for individuals with ID: preliminary data. *J. Intellect. Disabil. Res.* 54, 573–577. doi: 10.1111/j.1365-2788.2010.01282.x
- Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: a rejoinder on the binning procedure. *Behav. Res. Methods* 49, 653–673. doi: 10.3758/s13428-016-0721-5
- Worhach, J., Boduch, M., Zhang, B., and Maski, K. (2021). Remote assessment of cognition in kids and adolescents with daytime sleepiness: a pilot study of feasibility and reliability. *MedRxiv: Preprint. Server. Health. Sci.* doi: 10.1101/2021.03.24.21254190
- Zanini, G. A. V., Miranda, M. C., Cogo-moreira, H., Nouri, A., Fernández, A. L., and Pompéia, S. (2021). An adaptable, open-access test battery to study the fractionation of executive-functions in diverse populations. *Front. Psychol.* 12:627219. doi: 10.3389/fpsyg.2021.627219

Conflict of Interest: The authors declare no commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Segura and Pompéia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Contactless Method for Measuring Full-Day, Naturalistic Motor Behavior Using Wearable Inertial Sensors

John M. Franchak*, Vanessa Scott and Chuan Luo

Perception, Action, and Development Laboratory, Department of Psychology, University of California, Riverside, Riverside, CA, United States

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Sarah Berger,
College of Staten Island, United States
Kaya de Barbaro,
University of Texas at Austin,
United States

*Correspondence:

John M. Franchak
franchak@ucr.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 27 April 2021

Accepted: 20 September 2021

Published: 22 October 2021

Citation:

Franchak JM, Scott V and Luo C
(2021) A Contactless Method for
Measuring Full-Day, Naturalistic Motor
Behavior Using Wearable Inertial
Sensors. *Front. Psychol.* 12:701343.
doi: 10.3389/fpsyg.2021.701343

How can researchers best measure infants' motor experiences in the home? Body position—whether infants are held, supine, prone, sitting, or upright—is an important developmental experience. However, the standard way of measuring infant body position, video recording by an experimenter in the home, can only capture short instances, may bias measurements, and conflicts with physical distancing guidelines resulting from the COVID-19 pandemic. Here, we introduce and validate an alternative method that uses machine learning algorithms to classify infants' body position from a set of wearable inertial sensors. A laboratory study of 15 infants demonstrated that the method was sufficiently accurate to measure individual differences in the time that infants spent in each body position. Two case studies showed the feasibility of applying this method to testing infants in the home using a contactless equipment drop-off procedure.

Keywords: motor development, posture, body position, wearable sensors, human activity recognition, machine learning

1. INTRODUCTION

Infants' increasing ability to transition into and maintain balance in different body positions is a hallmark of the first year (Adolph and Franchak, 2017). At birth, newborns can only lay supine on their backs or prone on their bellies. Otherwise, they rely on caregivers to place them in different positions or hold them in their arms. With age, infants master the ability to sit independently, crawl in a prone position, stand upright, and walk. In this paper, we describe a new method to characterize infants' body positions—held by caregivers, supine, prone, sitting, and upright—across an entire day using machine learning classification of wearable inertial motion sensors. We begin by describing the importance of understanding infant body position and then review existing measurement approaches. Afterwards, we present two studies: A laboratory validation study that shows how wearable sensors can be used to accurately categorize infant body position, and case studies that demonstrate how feasibly the method can be adapted to collect data from infants in the home while relying on caregivers to administer the procedure.

Growing evidence suggests that acquiring more advanced control over body position augments infants' opportunities for learning and exploration (Gibson, 1988; Libertus and Hauf, 2017; Franchak, 2020). For example, infants' visual experiences differ according to body position: While prone, infants' field of view is dominated by the ground surface and objects near the body, whereas upright infants have a more expansive view of their surroundings that includes distant objects and faces (Franchak et al., 2011, 2018; Kretch et al., 2014; Luo and Franchak, 2020). Sitting facilitates visual and manual exploration of objects compared with laying prone or supine

(Soska and Adolph, 2014; Luo and Franchak, 2020). Upright locomotion (walking) compared with prone locomotion (crawling) allows infants to travel farther, more easily carry objects, and elicits different social responses from caregivers (Gibson, 1988; Adolph and Tamis-LeMonda, 2014; Karasik et al., 2014). Accordingly, learning to sit and walk is linked with downstream improvements in language learning and spatial cognition (Soska et al., 2010; Oudgenoeg-Paz et al., 2012, 2015; Walle and Campos, 2014; He et al., 2015; Walle, 2016; West et al., 2019, c.f. Moore et al., 2019). Presumably, these facilitative effects result from infants spending more time sitting, standing, and walking. For example, mastering the ability to sit independently nearly doubled the amount of time that 6-month-olds spent sitting (both independent and supported sitting) in daily life compared with 6-month-old non-sitters (Franchak, 2019). Infants who spend more time sitting have increased opportunities to explore objects. Yet, little data are available to describe how infants spend their time in different body positions across a typical day, and how the prevalence of different body positions changes with age and motor ability.

Video observation is the gold standard for measuring body position. But, video observation comes with several costs, especially with respect to the goal of describing natural, home experiences across a full day. Whereas, language researchers have profitably used day-long audio recordings to characterize the everyday language experiences of infants (e.g., Weisleder and Fernald, 2013; Bergelson et al., 2019), motor researchers have been limited to scoring body position recorded in relatively short (15–60 min) video observations (Karasik et al., 2011, 2015; Nickel et al., 2013; Thurman and Corbetta, 2017; Franchak et al., 2018). Although infants can wear an audio recorder that travels wherever they go, capturing infants' movements requires an experimenter to follow the infant from place to place while operating a camcorder. Furthermore, the presence of the experimenter in the home may lead to reactivity—altering infants' and caregivers' behaviors when observed (Tamis-LeMonda et al., 2017; Bergelson et al., 2019)—which threatens generalizability. Another threat to external validity is how time is sampled: A short visit from an experimenter scheduled at a convenient time is unlikely to be representative of the full spectrum of daily activities (e.g., nap routines, meal times, play, and errands) that may moderate motor behavior (Fausey et al., 2015; de Barbaro and Fausey, 2021; Kadooka et al., 2021). Other limitations of video observation are practical rather than scientific. Video recording an infant for an entire hour is laborious; to do so for an entire day would not be feasible. Even if it were possible to capture full day video recordings of an infant, frame-by-frame coding of body position would be a gargantuan task—slow but feasible in a small sample, but intractable at a larger scale—and storage of large, full-day video files creates a nontrivial data management challenge. As with audio, collecting video data in the home across an entire day presents challenges for maintaining participant privacy (Cychoz et al., 2020). Finally, physical distancing guidelines during the COVID-19 pandemic mean that an experimenter may not be permitted in the home to operate a video camera.

One alternative is to employ survey methods in lieu of direct observation. Surveys can be conducted remotely without an experimenter present in the home, addressing some limitations of video observation (i.e., reactivity, privacy, labor, data storage). Although retrospective diaries have been used to estimate infant body position and motor activity (Majnemer and Barr, 2005; Hnatiuk et al., 2013), their accuracy and reliability are questionable. For example, Majnemer and Barr (2005) asked caregivers to fill out a diary every 2–3 h to indicate the infants' position for each 5-min interval since the last entry. However, by 12 months of age infants change position an average of 2–4 times per minute when playing (Nickel et al., 2013; Thurman and Corbetta, 2017). Thus, it seems unlikely that a caregiver could accurately estimate the time spent in body positions using a retrospective diary. Ecological momentary assessment (EMA) is one alternative: Sending text message surveys to ask caregivers to report on infants' instantaneous body position every 1–2 h across the day provides a sparse, but accurate report (Franchak, 2019; Kadooka et al., 2021). Although this method may better capture full-day experiences compared with short video observation (and more accurately compared with retrospective diaries), it lacks the real-time position data that are provided by video coding.

Classifying body position from wearable sensors provides a third option that addresses the limitations of both video and survey methods. Lightweight inertial movement units (IMUs)—small sensors that contain an accelerometer and gyroscope—can be worn for the entire day or multiple days taped to the skin, embedded in clothing, or worn on a wristwatch (Cliff et al., 2009; de Barbaro, 2019; Lobo et al., 2019; Bruijns et al., 2020). Notably, an experimenter does not need to be present, and data can be recorded at a dense sampling rate in real time. Although video data must be collected and coded to train the classifier, the video-recorded portion can be brief (addressing privacy, data storage, and data coding labor concerns) while still providing a full-day measure of activity. Previous validation studies show that wearing lightweight sensors does not alter movements even in young infants (Jiang et al., 2018). Child and adult studies have successfully used wearable motion sensors to characterize the intensity of physical activity (e.g., sedentary vs. moderate-to-vigorous) using either cut points that set thresholds for different activity levels (Trost et al., 2012; Kuzik et al., 2015; Hager et al., 2017; Armstrong et al., 2019) or by training machine-learning algorithms to classify activity into different levels (Hagenbuchner et al., 2015; Trost et al., 2018).

Body position may be a more challenging behavior to classify compared with physical activity intensity. For example, an infant can be stationary or moving quickly while upright, suggesting that simple cut points or thresholds may not be suitable (Kwon et al., 2019). However, results from previous studies using machine learning to classify activity type in adults (Preece et al., 2009; Arif and Kattan, 2015) and children (Nam and Park, 2013; Zhao et al., 2013; Ren et al., 2016; Stewart et al., 2018) are encouraging. For example, Nam and Park (2013) used a support vector machine classifier to distinguish 11 activity types—including rolling, standing still, walking, crawling, and climbing—in a laboratory study of 16- to 29-month-olds. The classification accuracy was high (98.4%), suggesting that machine

learning classification of wearable sensors may be sufficiently sensitive to differentiate the activities of young children.

Despite an abundance of work with children and adults, only a handful of studies have investigated infants. A number of studies have used sensors worn on the wrists or ankles to estimate the frequency of limb movements in typical and atypical development (Smith et al., 2017; Jiang et al., 2018). Hewitt et al. (2019) used commercially-available sensors to detect one type of body position, prone, to estimate caregivers' adherence to "Tummy Time" recommendations. Greenspan et al. (2021) estimated body position angle using pitch angle cut-points from a single sensor embedded in a garment in 3-month-olds. Yao et al. (2019) used a pair of sensors, one worn by the infant and one worn by the caregiver, to train machine learning models that were able to accurately classify the time infants spent held by caregivers. Notably, the Yao et al. study validated their method "in the wild" by collecting data in the home rather than relying only on a laboratory sample, which suggests the feasibility of this method for our proposed application. Finally, one previous study measured body position in 7-month-old infants using a set of 4 IMUs embedded in a garment (Airaksinen et al., 2020). With all 4 sensors (accuracy declined using a single sensor or a pair of sensors), researchers were able to distinguish between supine, side-lying, and prone positions with 98% accuracy using a machine learning model.

Although recent work provides an encouraging outlook for measuring body position in infants (Yao et al., 2019; Airaksinen et al., 2020; Greenspan et al., 2021), there are several open questions. First, because past studies of body position (Airaksinen et al., 2020; Greenspan et al., 2021) did not include caregivers holding infants as a category, it is unknown whether our proposed body position categories—prone, supine, sitting, upright, and held by caregiver—can be accurately classified. Held by caregivers is critical because infants' bodies may seem to be configured in a similar way to another position while held (e.g., a caregiver cradling an infant might be in a similar body position to when they are supine in a crib or on the floor). For this reason, angle cut-points like those used in past work (Greenspan et al., 2021) are unlikely to capture differences in the five positions we aim to classify. Unless we can accurately distinguish when infants are held, it would not be possible to account for their body position across the day because infants are held as much as 50% of the time in a typical day (as measured using EMA, Franchak, 2019). Although the Yao et al. study measured caregiver holding time (but not other body positions), they used a pair of sensors (one worn by the infant and one worn by the caregiver). It is unclear whether sensors worn only by infants would be able to detect when they are held. Second, the Airaksinen et al. study's categories included sitting, however, sitting in daily life can take many forms—sitting on a caregiver's lap, sitting in a restrained seat, or sitting independently on the floor—that may make it harder to detect in the wild. In the current study, we trained and tested sitting in a variety of forms to be sure that we can capture the variability we expect to find across a full day in the home. Third, although a benefit of classifying behavior from wearable sensors is that an experimenter does not need to be present for the entire day, the classifiers still need to be trained on a set of

manually-coded ground truth data (e.g., body positions coded from video synchronized with sensor data). Given the regulatory issues arising from the COVID-19 pandemic, such as physical distancing and sanitation, we investigated the feasibility of using a stationary camera and sensors dropped off at participants' doorstep for training and validating a classifier without the researcher entering the home. But, it remains an open question whether an experimenter can remotely guide caregivers through the complex procedure of applying the sensors, synchronizing the sensors to the camera, and eliciting different body positions in view of the camera.

A remote drop-off procedure would have utility aside from addressing the immediate concerns of the COVID-19 pandemic. For families who feel uncomfortable with an experimenter visiting their homes, a remote drop-off provides a way to collect observational data without an experimenter's presence. Removing the need for an experimenter to spend an hour in the home—simply to pan a video camera—also reduces the experimenter's labor for collecting data. Most importantly, removing the experimenter's presence from the home—and the need to record video for long periods of time—can reduce reactivity. Indeed, caregivers spoke more to infants when video-recorded by a stationary camera than during an audio-only recording (Bergelson et al., 2019). Although our method uses a stationary camera, it is only needed for a brief video-recorded period followed by a full-day motion measurement (without video or experimenter presence). This will allow unobtrusive capture of behavior across a sufficiently long period to examine within-day variability of behavior (de Barbaro and Fausey, 2021) with minimum reactivity. Such data are crucial for testing the links between everyday experiences and subsequent development (Franchak, 2020). For example, one potential mechanism to explain why the acquisition of independent walking predicts increases in vocabulary development (Walle and Campos, 2014; Oudgenoeg-Paz et al., 2015) is that caregivers provide different language input to infants when infants are crawling compared with when they are walking (Karasik et al., 2014). However, since this difference was observed through experimenter-recorded video in the home, it is unknown how it generalizes across the day or whether such a difference persists when the experimenter and video camera are absent. Simultaneously recording speech with an audio recorder synchronized with body classification from motion sensors would provide full-day, unobtrusively-collected data to bear on this question.

2. LABORATORY STUDY: VALIDATING THE BODY POSITION CLASSIFICATION METHOD

The goal of the laboratory study was to test whether mutually-exclusive body position categories suitable for full-day testing—held by caregivers, supine, prone, sitting, and upright—could be accurately classified from infant-worn inertial sensors. We collected synchronized video and inertial sensor data while infants were in different body positions, and used those data to train classifiers and then validate them against the gold standard

(human coding from video observation). As in past work (Nam and Park, 2013; Yao et al., 2019; Airaksinen et al., 2020), our aim was to determine whether the overall accuracy of classification was high ($> 90\%$ of agreement between model predictions and ground truth data). Moreover, we assessed whether the method could accurately detect individual differences in how much time infants spend in different body positions, which is relevant for characterizing everyday motor experiences and their potential downstream effects on other areas of development (e.g., Soska et al., 2010; Oudgenoeg-Paz et al., 2012; Walle and Campos, 2014).

In order to identify the most accurate method for classifying body position, we compared two modeling techniques: *individual models* that were trained on each individual's data vs. *group models* that used a single model trained on all but one of the participants. Group models are more commonly used in activity recognition studies (e.g., Nam and Park, 2013; Yao et al., 2019; Airaksinen et al., 2020), and have several practical benefits, such as reducing complexity (only needing to train/tune a single model) and providing a generalizable method (group models can be used to classify data in participants for whom no ground truth training data were collected). We reasoned that although individual models take more work to create, they might lead to better accuracy in our use case for several reasons. First, individual models eliminated the possibility that variability in sensor placement across infants could add noise to the data. Second, given the wide range of ages (6–18 months), it allowed us to tailor models to the motor abilities of each infant. For example, the upright category could be dropped for the youngest infants who were never standing or walking. Moreover, the biomechanics of sitting likely differ between a 6-month-old and an 18-month-old, which could result in different motion features. Third, training and validating a model for each infant allows researchers to individually verify the data quality for each infant included in the analyses.

2.1. Materials and Methods

2.1.1. Participants

Participants were recruited from social media advertisements and local community recruitment events. The final sample consisted of 15 infants between 6 and 18 months of age (7 male, 8 female, M age = 11.28 months). Caregivers reported the ethnicity of infants as Hispanic/Latinx (9) or not Hispanic/Latinx (6). Caregivers reported the race of infants as White (10), More than One Race (2), Asian (1), and Other (1); one caregiver chose not to answer. An additional 7 infants were run in the study but could not be analyzed because of problems with the sensors (one or more sensors failed to record or stream data). Two additional infants were run in the study but excluded due to video recording failures, and one additional infant started the study but did not complete the session due to fussiness. Caregivers were compensated \$10 and given a children's book for their infant. The study was reviewed and approved by Institutional Review Board of the University California, Riverside. Caregivers provided their written informed consent to participate in this study and gave permission to record video and audio for both themselves and their infant before the study began.

2.1.2. Materials

Three MetaMotionR (Mbleintlab) inertial movement units (IMUs) were placed at the right hip, thigh, and ankle of infants and recorded accelerometer and gyroscope data at 50 Hz. Due to the high rate of sensor failures resulting in participant exclusion, we do not recommend use of this sensor and chose a different sensor for our subsequent projects. The IMU worn on the hip sat inside a clip fastened at the top of the infant's pant leg or diaper on the right side. The other two IMUs were placed in the pockets of Velcro bands strapped to the infant's right thigh (just above the knee) and right ankle. During the study, the IMUs streamed data via Bluetooth to a Raspberry Pi computer running Metabase software (Mbleintlab). A camcorder (Sony HDRCX330) held by an experimenter recorded infants' movements throughout the study so that body position could be coded later from video.

2.1.3. Procedure

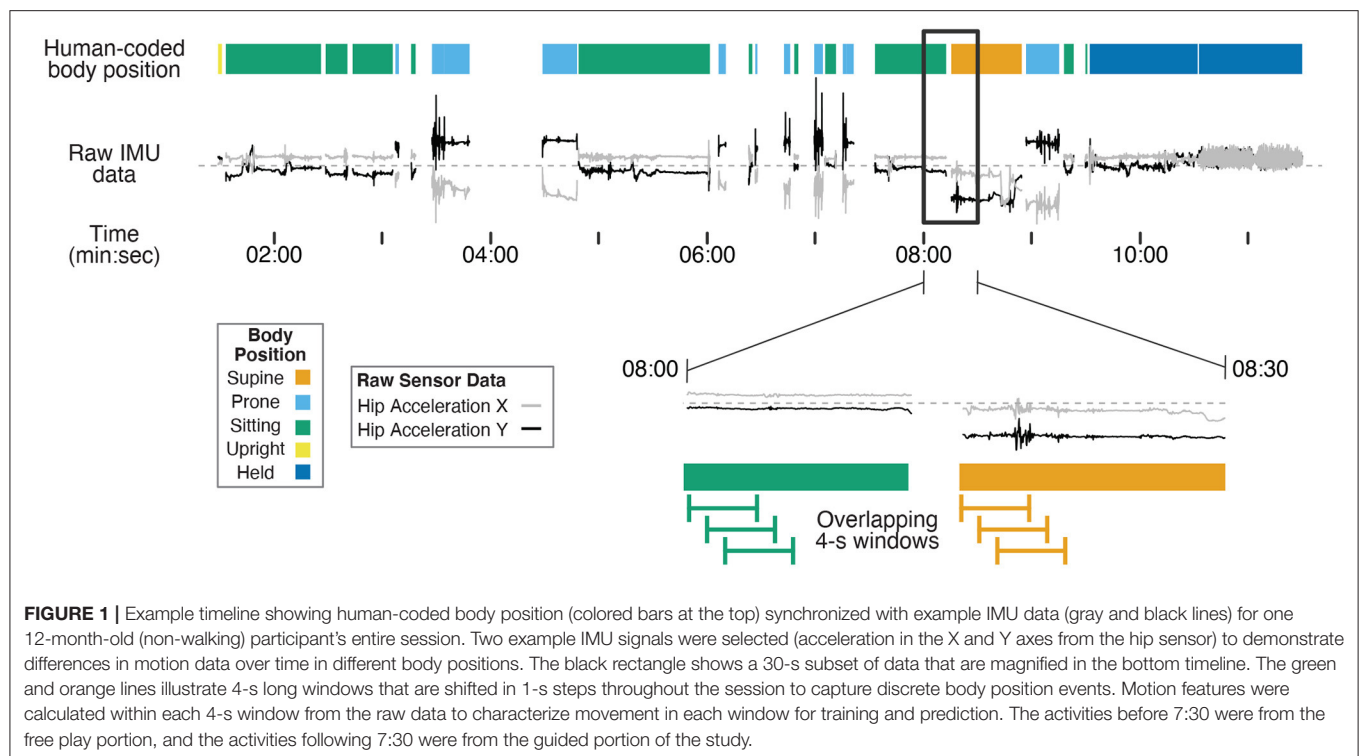
The study started with synchronizing the three IMUs to the video. To create an identifiable synchronization event in the motion tracking data, an experimenter raised all three sensors together and struck them against a surface in view of the camcorder with both the camcorder and sensors recording. After the synchronization event, the experimenter attached the three IMUs to the infant. The experimenter ensured the correct orientation of the IMUs by checking the arrow indicator on each IMU which faced forward toward the anterior plane with respect to the infant's body position.

After placing the IMUs on the infant, the experimenter guided the caregiver to put the infant in the following positions (assisted or non-assisted): standing upright, walking, crawling, sitting on the floor, lying supine, lying prone, held by a stationary caregiver, held by caregiver walking in place, and sitting restrained in a highchair. Each position lasted 1 min, and the total guided activities lasted approximately 10 min. After the guided activities, the caregivers were asked to play with their infants freely with toys for 5 min. During the free play portion, infants were permitted to move however they wished so that we could record spontaneous body positions. For some infants, the free play portion preceded the guided activities if the infant was fussy or resistant to the guided activities. An assistant held the camcorder and followed the infants throughout the guided and free-play activities to make sure the infant's body was always in view. To check synchronization, a second synchronization event was captured at the end of the study before turning off the video and IMU recordings.

2.1.4. Human Coding of Body Position

Human coders went through the third-person view videos recorded by the camcorder and identified infants' position in each frame using Datavyu software (www.datavyu.org). Body positions were identified as *supine*, *prone*, *sitting*, *upright*, or *held by caregiver*. **Figure 1** shows an example timeline of position codes over the session for one infant.

Supine was coded when the infant was lying on their back. Prone was coded when the infant was lying flat on the stomach or in a crawling position (either stationary or locomoting). Sitting was coded when the infant was sitting on a surface (e.g., a



couch or floor, with or without support from the caregiver), the highchair, or on the caregiver's lap. Upright was coded when infants were standing, walking, or cruising along furniture. Held by caregiver was coded when the infant was carried in the caregivers' arms off the ground, excluding times that they were seated on the caregiver's lap. Positions that could not be identified as any of these categories (such as times in transition between body positions) or times where the sensors were briefly removed/adjusted were excluded from coding (i.e., gaps between data in **Figure 1**). Each video was coded in its entirety by two coders. The interrater reliability between the two coders was high across the 15 videos (overall agreement = 97.6%, $kappa = 0.966$).

2.1.5. Machine-Learning Classification of Body Position

The data were processed in three steps. First, the timeseries of accelerometer and gyroscope data were synchronized to the human-coded body position events. Second, we applied a moving window to the synchronized timeseries to create 4-s long events, and extracted motion features that characterized each event. Finally, we trained random forest classifiers (both individual models and group models) to predict the body position categories for each participant based on the motion features in the 4-s windows.

2.1.5.1. Synchronization

A researcher plotted the accelerometer time series in Matlab and identified the timestamp that corresponded to the acceleration peak at the moment the sensors were struck during the synchronization event. That timestamp was subtracted from the

other timestamps to define the synchronization event as time 0. Likewise, Datavyu video coding software was used to find the moment the sensors were struck against the surface in the video, and that time was defined as time 0 for body position codes. In doing so, human-coded body position was synchronized with the motion data. The synchronization event at the end of the session was used to confirm that the synchronization was correct and that no drift correction was needed. The onsets and offsets of each human-coded body position were used to construct a 50 Hz time series of body position categories, providing a body position code that corresponded to each sample of motion data.

2.1.5.2. Window Creation and Feature Generation

As in previous studies in human activity recognition (Preece et al., 2009; Nam and Park, 2013; Airaksinen et al., 2020), overlapping moving windows were applied to the synchronized motion and body position timeseries in Matlab: 4-s windows were extracted every 1 s from the first synchronization point to the end of the session. The magnified timeline at the bottom of **Figure 1** shows examples of the overlapping 4-s windows. As such, each 4-s window contained 200 samples of 50 Hz motion data. We omitted any window during which a position category was present for less than 3 s of the 4-s window to avoid analyzing windows that included transition movements between positions or a mix of two different body positions.

Across the 200 samples in a window, we calculated 10 summary statistics—the mean, standard deviation, skew, kurtosis, minimum, median, maximum, 25th percentile, 75th percentile, and sum—for each combination of 3 sensor locations (ankle, thigh, and hip), 2 sensor signals (acceleration, gyroscope),

and 3 axes (X, Y, Z for acceleration; roll, pitch, yaw for gyroscope). For example, 10 summary statistics described the ankle's acceleration in the Z dimension. In total, 10 statistics \times 3 sensor locations \times 2 sensor signals \times 3 axes resulted in 180 features. In addition, we calculated the sum and magnitude of movement in each axis across the three sensor locations and the sum and magnitude of movement across axes within each sensor. Finally, we calculated correlations and difference scores between each pair of axes within a sensor and between each pair of sensors for a given axis. These cross-sensor and cross-axis features brought the motion feature total to 204.

2.1.5.3. Model Training

To train and validate *individual models*, each participant's data were separated into a training set that was used to train the model, and a testing set that was held out for validation. In order to mimic the intended use of this method—using video coded at the start of the day to train a model for predicting body position over the rest of the day, we used the first 60% of each participant's data as the training set and the remaining 40% as the testing set. However, because of the sequential nature of our guided activities, selecting the first 60% chronologically would include some activities and exclude others. Thus, we selected the first 60% of data *within each body position category* for the training set to ensure that there were sufficient data to train the models on all positions. To train and validate *group models*, we used a leave-one-out cross-validation technique. A group model was trained using all of the data from 14/15 participants, and then the remaining participant's data served as the testing set. In this way, we could report classification accuracy for each participant (as predicted from a model trained on all other participants). As in Airaksinen et al. (2020) we excluded windows in which the primary and reliability coders disagreed to ensure that only unambiguous events were used in training across both types of models.

Machine learning models were trained in R using the *randomForest* package to create random forest classifiers (Liaw and Wiener, 2002). The random forest algorithm (Breiman, 2001) uses an ensemble of many decision trees—each trained on a random subset of motion features and a random subset of the training data—to avoid overfitting and improve generalization to new cases (Strobl et al., 2009). Prior work shows random forests are well-suited to classifying motor activity (Trost et al., 2018; Yao et al., 2019). By training hundreds of trees on different subsets of features, the classifier detects which features (of our set of 204) are most useful in classifying the categories we chose. In a preliminary step, we optimized two parameters, the number of trees and the “mtry” parameter, by training and testing classification accuracy across a range of parameter values. The optimal number of trees trained in the model was 750 (using more trees took longer processing time without significant gains in model accuracy). The “mtry” parameter refers to how many features are randomly selected in each tree, and the default value was optimal (square root of total number of features). Regardless, performance varied little depending on the values of these parameters. Using the optimal parameters, a random forest model was created based on each participant's training data

TABLE 1 | Unweighted, overall accuracy, and Cohen's Kappa for each individual participant in the lab validation study.

| Individual | | Group | |
|------------|-------|----------|-------|
| Accuracy | Kappa | Accuracy | Kappa |
| 0.92 | 0.91 | 0.95 | 0.94 |
| 0.94 | 0.90 | 0.95 | 0.91 |
| 0.96 | 0.94 | 0.84 | 0.56 |
| 0.96 | 0.96 | 0.82 | 0.82 |
| 0.97 | 0.70 | 0.94 | 0.81 |
| 0.97 | 0.94 | 0.99 | 0.79 |
| 0.98 | 0.94 | 0.90 | 0.60 |
| 0.99 | 0.97 | 1.00 | 1.00 |
| 0.99 | 0.98 | 0.98 | 0.95 |
| 0.99 | 0.99 | 0.99 | 0.98 |
| 1.00 | 1.00 | 0.89 | 0.84 |
| 1.00 | 1.00 | 0.91 | 0.75 |
| 1.00 | 1.00 | 0.95 | 0.73 |
| 1.00 | 1.00 | 0.92 | 0.88 |
| 1.00 | 1.00 | 0.94 | 0.78 |
| 0.98 | 0.95 | 0.93 | 0.82 |

Accuracy is reported separately for individual vs. grouped models. Bottom row shows average overall accuracy and Kappa values across participants.

(individual model) and from all but one participants' data (group model). The *predict* function was then used to apply the model to the motion features in the testing data set to classify each window, and provide a set of predicted categories to compare to the human-coded categories. For individual models, the testing set was the 40% of data held for testing; for group models, the testing set was the “left out” participant. In both cases, testing data were independent from data used to train the model that was validated and included behavior from both the guided activities and the free play portion.

2.2. Results

To validate models, we compared the classifier prediction to the ground truth (human-coded body position categories) for each window in the testing data set. The overall accuracy (across body position categories) for each participant was calculated as the percentage of windows in which the model prediction matched the human-coded position. Because windows were of equal length (4 s), accuracy can likewise be interpreted as the percentage of time that was correctly predicted by the model. **Table 1** shows the accuracy for each participant for the individual and the group models. For individual models, overall accuracy averaged $M = 97.9\%$ ($SD = 2.37\%$, ranging from a minimum of 92.4% to a maximum of 100%), similar to or exceeding the accuracy reported in related investigations (Nam and Park, 2013; Yao et al., 2019; Airaksinen et al., 2020). For group models, overall accuracy was lower ($M = 93.2\%$, $SD = 0.053$), but still strong. A paired samples *t*-test confirmed that individual models yielded superior accuracy, $t(14) = -3.28$, $p = 0.0055$.

Although the overall accuracy was excellent, it can overestimate the performance of the model if it does better at predicting more prevalent categories (e.g., sitting) and misses less prevalent categories (e.g., prone). Despite attempting to elicit each body position for a set amount of time for each infant during the guided session, not all infants exhibited each behavior (e.g., infants who could roll might refuse to remain supine and/or prone). Every infant sat and every infant was held by a caregiver, but the prevalence varied greatly across infants of different ages and motor abilities. **Figure 2** and **Table 2** show the mean prevalence (% of session spent in each position). Infants spent the most time sitting ($M = 45.98\%$, $0.9\text{--}70.2\%$) followed by held ($M = 33.75\%$, $21.6\text{--}84.8\%$). Upright positions were recorded in 10/15 infants with an average of $M = 16.02\%$ (out of infants who were upright), and ranged from a minimum of 2.8% to a maximum of 55.5% of the session. Supine (9/15 infants) and prone (11/15 infants) were observed least often. Infants were supine $M = 8.61\%$ of the time ($1.8\text{--}17.7\%$) and were prone $M = 6.04\%$ of the time ($0.4\text{--}13.1\%$).

To account for differences in prevalence, we calculated Cohen's Kappa, a measurement of agreement for classification data that controls for the base rate of different classes.

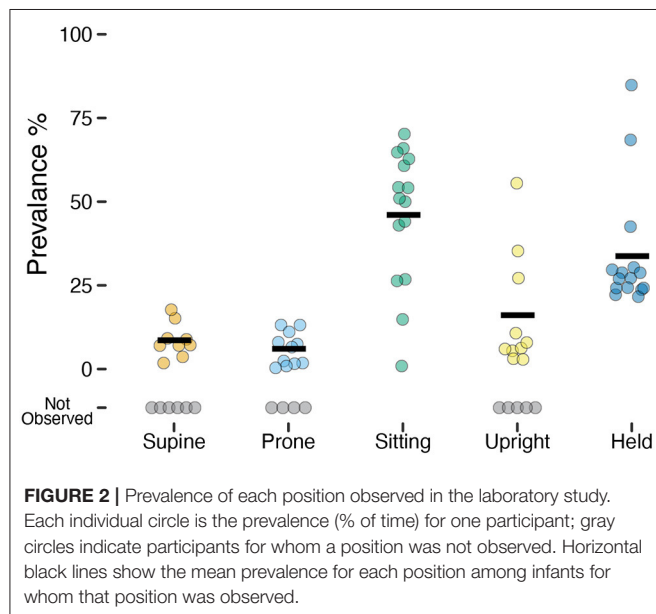


Table 1 shows the Kappa values for each participant, which were significantly higher on average for the individual models ($M = 0.95$, $SD = 0.076$) compared with the group models [$M = 0.82$, $SD = 0.129$, $t(14) = -3.36$, $p = 0.0047$]. As in past work (Greenspan et al., 2021), we interpreted the Kappa values according to Landis and Koch (1977) ranges: 0.81–1.00 “Almost Perfect,” 0.61–0.80 “Substantial,” 0.41–0.60 “Moderate,” 0.21–0.40 “Fair,” 0–0.20 “Slight to Poor.” Based on those guidelines, 14/15 participants’ classifications from the individual models were Almost Perfect and 1/15 was Substantial. In contrast, 9/15 participants’ classifications from the group models were Almost Perfect, 4/5 were Substantial, and 2/15 were Moderate. Given the better performance of individual models, across both accuracy metrics (overall accuracy and Kappa), we opted to use individual models (and focus solely on those models for the remaining results).

2.2.1. Sensitivity and Positive Predictive Value by Body Position

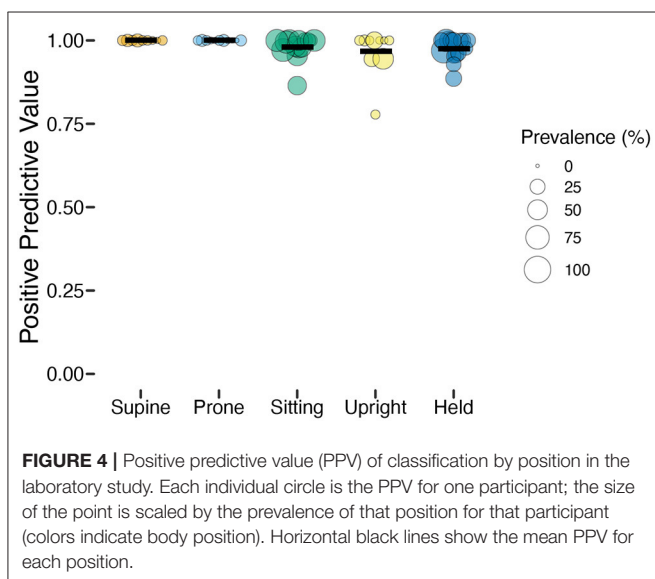
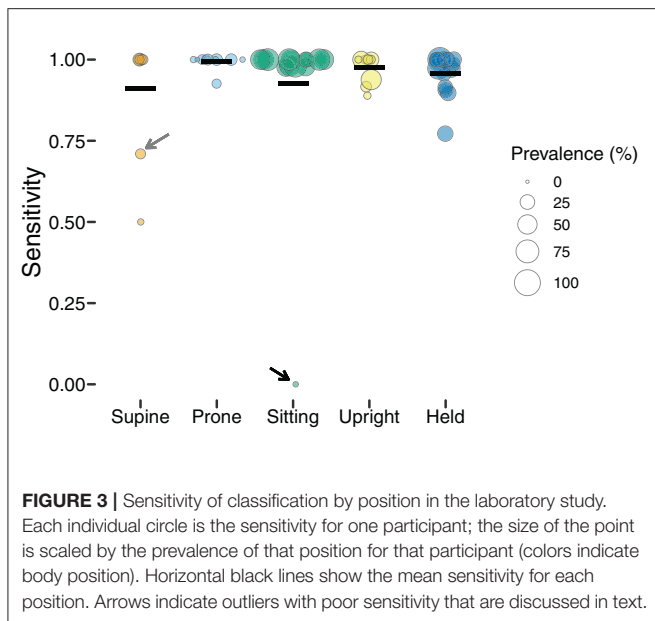
To better understand the classification performance within each body position, we calculated the *sensitivity* (the proportion of actual occurrences of each body position that were correctly predicted; also referred to as recall) and the *positive predictive value* (the proportion of predictions for a given category that corresponded to actual occurrences; also referred to as precision). **Table 2** summarizes sensitivity and positive prediction value (PPV) by position using the individual models.

Figure 3 shows the sensitivity of classifications by body position, and each individual point shows one participant’s data (size is scaled to the prevalence of the position, with larger symbols indicating greater frequency). Although mean sensitivity was generally high ($M_s > 0.91$), there was variability among participants and positions. For example, one infant’s supine sensitivity was 0.71 (indicated by the gray arrow), indicating that of the 31 actual supine 4-s windows, the model only predicted 22 supine windows. The worst outlier was one infant’s sitting position that had a sensitivity of 0 (indicated by the black arrow). Possibly, sensitivity related to prevalence. For that infant, there were only 2 windows in the testing dataset to classify and both were missed. Because training datasets were similarly limited by the number of windows containing sitting, there were likely insufficient data to train the sitting category for that infant.

Whereas, sensitivity varied among individuals and positions, positive prediction value (PPV) was uniformly high (**Table 2**).

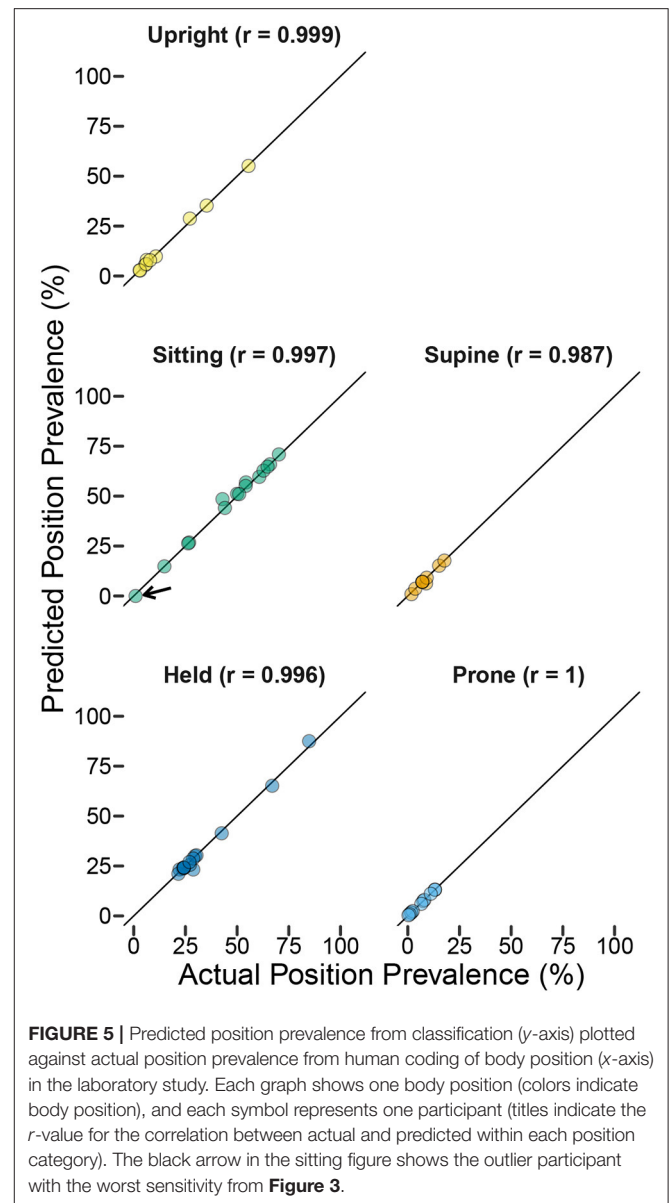
TABLE 2 | Prevalence, sensitivity, and positive predictive value by body position for the lab validation study testing dataset.

| Position | Prevalence | Sensitivity | | | | Positive predictive value | | | |
|----------|------------|-------------|-----------|-------|-------|---------------------------|-----------|-------|-------|
| | | <i>M</i> | <i>SD</i> | Min | Max | <i>M</i> | <i>SD</i> | Min | Max |
| Supine | 8.61 | 0.912 | 0.182 | 0.500 | 1.000 | 1.000 | 0.000 | 1.000 | 1.000 |
| Prone | 6.04 | 0.993 | 0.022 | 0.926 | 1.000 | 1.000 | 0.000 | 1.000 | 1.000 |
| Sitting | 45.98 | 0.928 | 0.257 | 0.000 | 1.000 | 0.981 | 0.036 | 0.865 | 1.000 |
| Upright | 16.02 | 0.974 | 0.043 | 0.889 | 1.000 | 0.967 | 0.070 | 0.778 | 1.000 |
| Held | 33.75 | 0.959 | 0.064 | 0.772 | 1.000 | 0.976 | 0.034 | 0.886 | 1.000 |



As **Figure 4** shows, upright had the worst average PPV ($M = 0.976$) and lowest minimum (0.778). For the participant with the lowest PPV, a value of 0.778 meant that of 9 detected upright windows, only 7 corresponded to actual upright behavior.

Overall, the high (> 0.90) average sensitivity and PPV within each class indicate that the classifiers performed well for each position despite varying prevalence. However, there were a few concerning individual outliers for sensitivity. Although outliers such as these might be addressed in future work by collecting and testing with a larger dataset, it is important to know what impact they might have on the interpretation of the data, and in particular, for revealing individual differences in position durations.



2.2.2. Capturing Individual Differences in Position Duration

The intended use of this method is to describe individual differences in the relative amounts of time that infants spend in different body positions. To what extent did the prediction of accumulated time spent in each position reflect the actual time spent in each position? We calculated each participant's predicted prevalence as the proportion of 4-s windows classified in each category divided by the total number of windows in their testing dataset. **Figure 5** shows scatterplots of actual vs. predicted prevalence for each position. Correlations (shown in the titles of each scatterplot) were very strong ($r_s > 0.987$), indicating excellent consistency between model classification and human coding in detecting individual differences in position prevalence. It is interesting to note that even the most extreme outlier for

sensitivity (sitting participant indicated by the black arrow whose sensitivity was 0) did not disrupt the correlation. Since outliers were for participants/positions with low prevalence, missing events (or even missing every event) still resulted in a good-enough predicted value for the purpose of capturing individual differences in posture duration between infants.

3. CASE STUDY: FEASIBILITY OF CONTACTLESS HOME DATA COLLECTION

The home data collection procedure described below addresses challenges we faced in adapting the laboratory protocol to measuring body position in the home during the COVID-19 pandemic. The risk of COVID-19 transmission between people in an indoor space, especially over prolonged periods of time, meant that the two experimenters could not enter the family's home to place the IMUs, guide the family through the procedures, and control the video camera. Instead, we developed a new, contactless protocol in which the experimenter dropped off equipment outside the family's door and guided the caregiver through procedures over the phone. However, relying on the caregiver to place the IMUs correctly, position the video camera to record infant behavior, and create synchronization events raises additional opportunities for error. Below, we detail several new procedures we developed to address those concerns: designing a customized pair of leggings with embedded IMUs to ensure the sensors are placed correctly by the caregiver, using a 360° camera to capture whole-room video even when camera placement is sub-optimal, and asking caregivers to record daily events that might disrupt IMU recording (i.e., diaper changes and naps).

Although the procedure is similar in many ways to the laboratory study, testing the new method on two case study participants helps to show whether it is feasible to collect high quality data despite major changes to how the procedure was implemented. Major differences between the laboratory study and the home data collection include: using a different set of IMU sensors embedded in a pair of leggings (rather than strapped to the infant), relying on caregivers to correctly place the leggings on the infant, using a fixed camera rather than an experimenter-operated camera to collect training/testing data, asking caregivers to elicit infant body positions and perform synchronization checks in view of the video camera, and collecting data over long periods of time (8 h of home data vs. 15 min of laboratory data). With the experimenter only able to communicate with the caregiver over the phone, any mistakes in equipment placement, synchronization, or body position tasks would not be caught by the experimenter until many hours later when the equipment was retrieved and the experimenter could check the video. As such, we report case study data from two participants to show the feasibility of collecting data (of sufficient quality to build body position classification models) after making these changes. Although we report classification accuracy for those two participants, validation data from a larger sample will be needed to determine if the method consistently allows for accurate body position classification.

3.1. Materials and Methods

3.1.1. Participants

Two participants, an 11-month-old infant (Participant A) and a 10.5-month-old infant (Participant B), were tested using the new contactless procedure. Neither infant could walk independently, but both could stand, cruise along furniture, and walk while supported with a push toy or caregivers' assistance.

3.1.2. Materials

To adapt the position classification method for testing in the home during the COVID-19 pandemic, data collection was conducted through a "guided drop-off" procedure. The caregiver received sanitized equipment in a sealed bucket left by the experimenter at their door. The bucket contained 4 Biostamp IMUs (MC10) embedded in a pair of customized infant leggings, a 360° camera on a tripod (Insta360 One R), sanitizing supplies, and paperwork.

The 4 IMUs were placed at the hip and ankle of the infants on both the right and left legs (testing from the lab study revealed that the thigh sensor was the least informative). The Biostamp IMUs are designed for full day recording: They have a long battery life (about 14 h) and record to onboard memory without the need to stream to a device or connect to the internet. Each IMU sensor recorded motion from an accelerometer and gyroscope at 62.5 Hz.

To minimize the possibility of caregivers placing the IMUs incorrectly on infants, a pair of customized leggings were fabricated with 4 small pockets sewn inside the hip and ankle positions of each leg. The snug, elastic fabric kept each sensor tight against the body so that they would not bounce or move independently from the body. The experimenter placed the sensors inside the garment before drop-off to ensure that sensors were oriented and labeled correctly (i.e., sensor A corresponded to the right hip location). The front and back of the garment were clearly labeled so that caregivers would put them on infants in the correct orientation.

We previously relied on an experimenter to operate a handheld camera so that the infant was always in view for body position coding. Without an experimenter in the home, the camera needed to be placed on a tripod. However, that could lead to sub-optimal views and high portions of the time where the infant is out of the video. To address this limitation, we used a camera that recorded in 360° (Insta360 One R). The caregivers were instructed to place the camera on a tabletop tripod in the room where their infants would spend the majority of the day, and were asked to move the camera if the infant left the room for an extended period of time. Since the camera simultaneously records in all directions, the placement of the camera in the room mattered less compared to using a traditional camera with a limited field of view (however, view of the infant could still be obstructed by furniture or people moving around in the room). After the study, the experimenter used specialized camera software to digitally orient the camera so that it exported a video with the infant in view at all times.

The paperwork included the consent form, instructions for how to set up the camera and put on the leggings, and a form that caregivers used to document times when the IMUs were

taken off the infants (e.g., diaper changes, naps, excursions out of the home).

3.1.3. Procedure

The procedure consists of a prior-day orientation call, a morning equipment drop-off, an experimenter-guided video session, and a sensor-only recording period for the rest of the day.

3.1.3.1. Prior Day Orientation Call

The participant was contacted a day before participation day to confirm their appointment. During this phone call the experimenter explained the contactless drop-off procedure, gave an overview of the equipment, and explained the consent form to prepare for the participation day.

3.1.3.2. Contactless Equipment Drop-Off

On the participation day, the experimenter brought the equipment bucket—containing sterilized, preconfigured equipment and paperwork—to the participant's home. Importantly, the IMUs were already set to record and were placed correctly within the leggings. When arriving at the participant's door, the experimenter started recording the 360° camera and created a synchronization point by striking the leggings (with the IMUs inside) in view of the camera. Afterwards, the experimenter went back to their vehicle and notified the participant over the phone that the equipment was ready to be picked up.

3.1.3.3. Guided Video Task

While on the phone with the experimenter, the caregiver was asked to open the bucket and then read and sign the consent form. Next, the caregiver was asked to place the 360° camera in an optimal location for video capture (e.g., a coffee table or TV stand). Then, the experimenter asked the caregiver to dress the infant in the leggings and provided prompts to check that the garment was worn correctly.

With all equipment recording, the experimenter (via phone) guided the caregiver through a set of procedures to elicit different body positions for training and testing the classification model. These tasks were the same as the laboratory tasks, but administered by the caregiver instead of the experimenter. The series of guided tasks involved the caregiver placing the infant in different positions: lying on their back (supine), lying on their stomach while stationary (prone), sitting on the floor (with support, if needed), crawling on the floor (if able), walking (if able, caregiver providing support if needed), standing still (if able, caregiver providing support if needed), picking up and holding child off the ground, sitting in a restrained seat (e.g., high chair). Each position lasted approximately 1 min.

Afterwards, the researcher asked the caregiver to create another synchronization event by removing the leggings from their infant, holding the leggings up in the air in view of the camera, and dropping them to the floor. Next, the caregiver was instructed to place the leggings back on their infant and spend 10 min playing with the infant in view of the camera. After receiving those instructions, the phone call with the experimenter ended.

3.1.3.4. Sensor-Only Recording and Material Pick-Up

After the 10 min of free play, the caregiver and infant went about their day as usual with the IMUs continuing to record for the next 8 h or until the experimenter had to pick up the equipment. The only responsibility for the caregivers during the rest of the day was to indicate every time they removed the leggings from the child for any reason (e.g., diaper changes, naps) on the paper log form. This allowed us to omit periods of the day during which the IMUs should not be analyzed.

The 360° camera continued to record until the battery ran out, so the caregiver was asked to position the camera in the room with the infant until the camera stopped recording. The camera could record 90–180 min depending on camera settings we used (in the second case study session we lowered the recording quality to increase recording time). However, because the experimenter started the camera recording before dropping off the equipment on the doorstep, the portion with the infant in view of the camera could vary substantially. For Participant A, the recording lasted 90 min with approximately 45 min of footage of the infant (there was a delay between dropping off the equipment and the camera recording the infant, and the infant went out of view toward the end of recording). For Participant B, we adjusted the settings to record a longer video (the recording lasted 180 min), and the infant was in view for almost the entire 180-min period.

Caregivers could call the experimenter during the day if they encountered any problems. The experimenter scheduled a time to pick up the equipment bucket from the participant's door in the evening or the following morning. All materials were then sterilized following CDC protocols in preparation for the next participant.

3.1.4. Video Processing and Coding

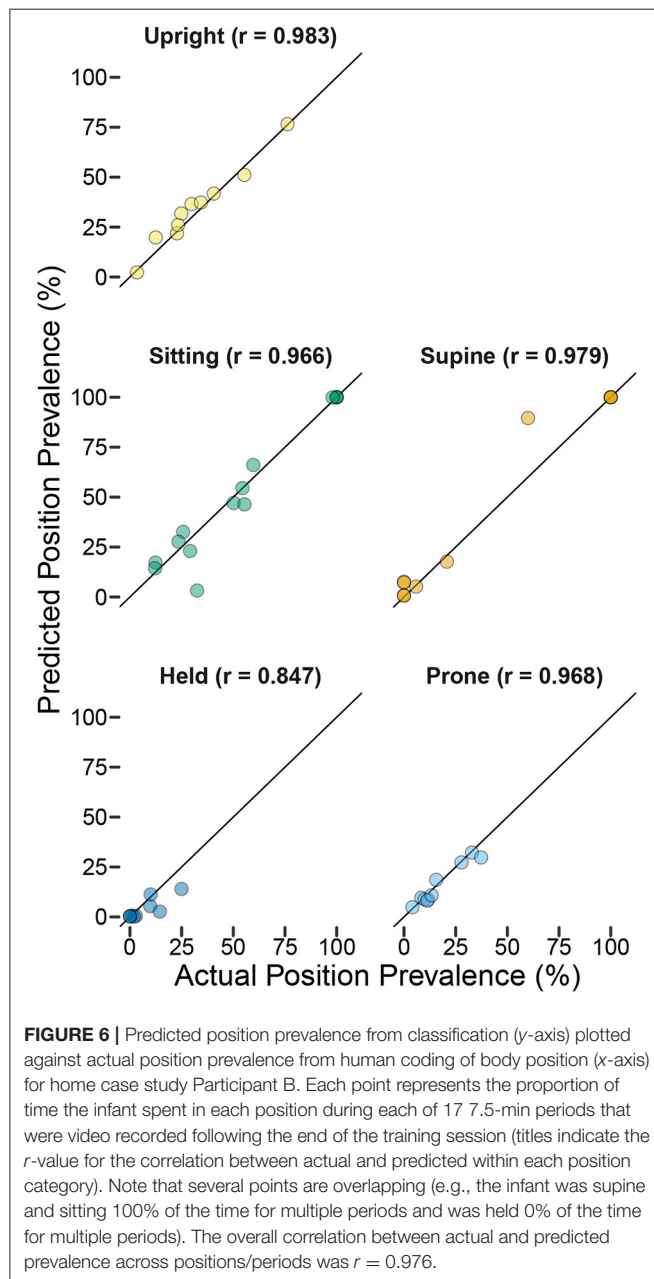
To prepare video data to be coded in Datavyu, an experimenter needed to manually edit the video footage to create a regular field of view video from the 360° video, which was in a proprietary format consisting of two hemispherical video files. Insta360 Studio software allowed the research to select a portion of the 360° video to bring into view. Camera orientations could be tagged at specific times, essentially allowing the researcher to pan the video camera—after the fact—to maintain the infant in view. After exporting a regular field of view video with the infant in view, the coders then identified the infant's position in each frame using the same coding categories as before: supine, prone, sitting, upright, or held by caregiver.

3.2. Case Study Results

Each participant's video was coded and synchronized with data from the 4 IMUs worn in the leggings. Data from the guided session (15 min of elicited body positions plus 10 min of free play) were combined and then divided into training and testing datasets. As before, individual models were created using the first 60% of each position type for training the model and the remaining 40% for testing. We compared the predicted positions from the random forest model to the actual coded positions in the testing data to assess the performance of the classifier. The overall accuracy was 85.2% for Participant A ($\kappa = 0.80$) and 86.6% for Participant B ($\kappa = 0.76$). **Table 3** shows

TABLE 3 | Prevalence, sensitivity, and positive predictive value (PPV) by body position for the testing datasets used to assess case studies (Participants A and B).

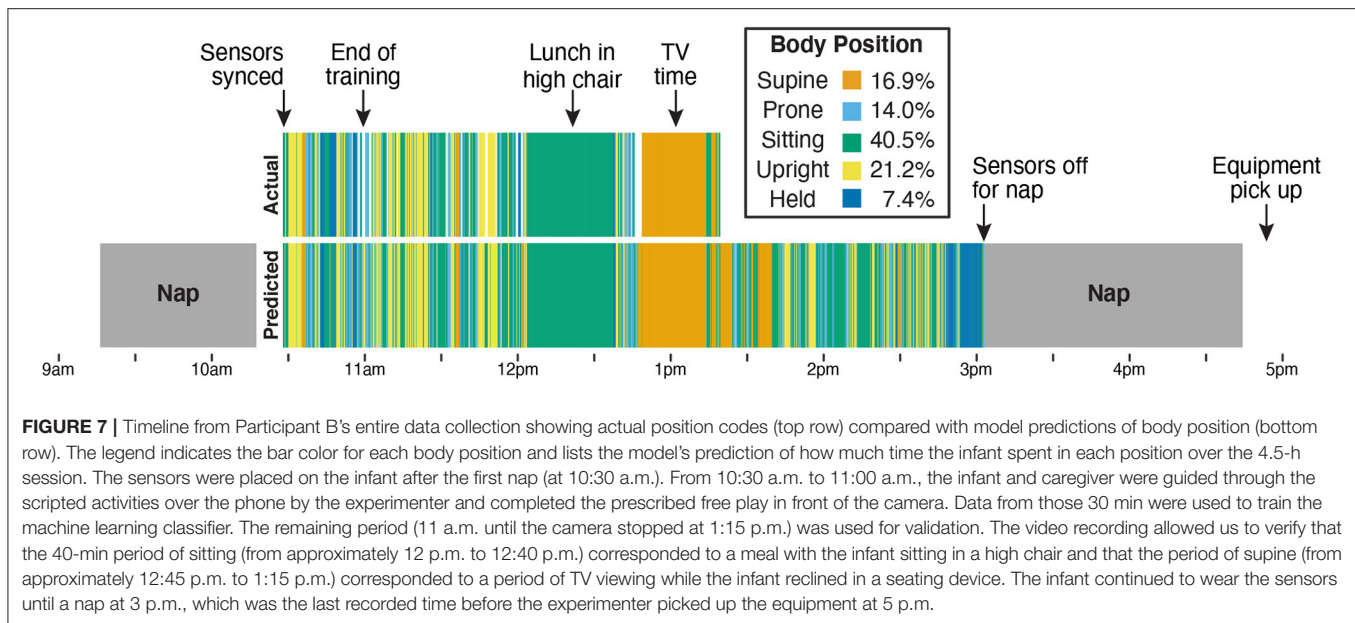
| Position | Participant A | | | Participant B | | |
|----------|---------------|-------------|-------|---------------|-------------|-------|
| | Prevalence | Sensitivity | PPV | Prevalence | Sensitivity | PPV |
| Supine | 6.91 | 1.000 | 1.000 | 22.74 | 0.973 | 0.877 |
| Prone | 10.22 | 0.676 | 0.833 | 10.14 | 0.671 | 0.728 |
| Sitting | 53.59 | 0.951 | 0.846 | 44.29 | 0.881 | 0.906 |
| Upright | 16.71 | 0.595 | 0.758 | 19.00 | 0.892 | 0.835 |
| Held | 12.57 | 0.846 | 0.928 | 3.83 | 0.453 | 0.837 |



the prevalence, sensitivity, and PPV for each of the five body positions for each participant. Overall, accuracy, Kappas, and sensitivity were weaker compared to the laboratory study, but still within acceptable levels (e.g., Yao et al., 2019; Greenspan et al., 2021).

As in the laboratory study, we found that the models performed well at detecting relative differences in the durations of different body positions even when sensitivity was less than ideal. To get a sense of differences in relative durations of positions over time within each infant, we used all available video that followed the guided tasks and free play (e.g., until the battery ran out or the infant was no longer on camera) to code the durations of every body position in 7.5-min intervals. For Participant A, 30 min of video were available (4 7.5-min periods), and for Participant B, 127.5 min of video were available (17 7.5-min periods). Within each period, we calculated the percentage of time in each body position predicted by the model compared to the actual percentage of time coded by hand. Correlations between actual vs. predicted percentages were strong: $r = 0.911$ across positions for Participant A and $r = 0.976$ for Participant B. Within-position scatterplots and correlations are shown in **Figure 6** for Participant B, for whom sufficient data were available. Although the correlations were weaker compared to the laboratory study, they suggest that these models can distinguish changes in the relative duration of different positions throughout the day.

Figure 7 shows a timeline of actual and predicted body positions during the entire recording session for Participant B, providing an example of the type of data afforded by this method. The sensors were synchronized and applied to the infant after her morning nap, and from 10:30 a.m. to 11:00 a.m. the infant and caregiver participated in the guided activities and the required free play portion that were used as training data. The next 2 h (until 1:15 p.m. when the camera battery ran out) were recorded on video and used to calculate the correlations in **Figure 6** and the validation statistics in **Table 3**. We were able to use the video to confirm two notable events in the timeline: A long period of sitting while the infant had lunch in a high chair, and a long period of supine while the infant watched TV in a rocking cradle. The sensors continued to record until the infant took a second nap at 3:00 p.m., and were picked up by the experimenter following the nap. The legend in **Figure 7** shows the proportion



of each body position predicted by the model across the entire sensor recording period.

4. DISCUSSION

The current studies demonstrate the validity and feasibility of classifying infant body positions from wearable inertial sensors. Moving beyond past work that classified only holding events (Yao et al., 2019) or body positions that omitted holding and upright as categories (Airaksinen et al., 2020), our laboratory study classified five body positions that be applied full-day behavior in the home, across activities that may include different forms of each body position (e.g., sitting on the floor during play, sitting in a high chair during a meal). Although sensitivity varied among participants and body positions, the classification system was able to reveal individual differences in time spent between different body positions between infants. The case studies went a step further to provide a proof-of-concept of how the method could be employed in the home across a long recording period. For both case study participants, we successfully collected video and motion data in the home by guiding caregivers through a contactless equipment drop-off procedure. The resulting body position classifiers—trained from data in which no experimenter entered the home or operated the equipment—were sufficiently accurate to measure intra-individual changes in body position over time, suggesting that the procedure could be carried out successfully by caregivers who received instructions over the phone.

Full-day recordings of body position have the potential to transform our understanding of everyday motor behavior in a similar way that wearable audio recorders have changed the study of language development. Wearable audio recorders capture the entire day (or even multiple days) of language input in the home (Weisleder and Fernald, 2013). The language input

infants receive differs between the lab and real life, depends on the activity context, and can be biased by the presence of an experimenter (Tamis-LeMonda et al., 2017; Bergelson et al., 2019). Moreover, recorders such as the LENA automatically score metrics about language input to reduce the need for laborious transcription. Although our method of body position classification still depends on collecting and scoring video data, a 30-min training period at the start of the day is enough to then turn off the cameras and unobtrusively record and classify body position for the remainder of the day (or in the future, multiple days).

As real-time, full-day motor experience data become available, what might we learn? Although Figure 7 shows “only” 8 h in the life of one infant, it is striking to observe the heterogeneity in motor activities across the day. The late morning and early afternoon were marked with frequent changes between different positions as the infant engaged in unrestrained play. In contrast, the lunch and TV times created long, interrupted bouts of a single body position. As more data become available from infants of different ages, motor abilities, and caregivers, we expect to see large inter- and intra-individual differences in body position. Indeed, our ongoing work using ecological momentary assessment to record infants' activities (e.g., play, feeding, media viewing, errands, etc.) shows that play is more frequent than any other activity for 11- to 13-month-olds (feeding is the second most prevalent), but play time differs greatly between infants (Kadooka et al., 2021). Some infants played for one third of the waking day, whereas others played for two thirds. Most likely, differences in daily activities provide a partial explanation for why body position rates measured in laboratory play (Thurman and Corbetta, 2017; Franchak et al., 2018) do not correspond to those measured in full-day EMA surveys (Franchak, 2019). Full-day timelines from wearable sensors will be even better suited to explain differences between the laboratory and the home because

they provide dense, real-time data (tens of thousands of samples a day) compared with the 8–10 total samples yielded through EMA notifications every hour.

Although the results of our validation and case studies are promising, there is still reason to be cautious as we apply the method to full-day testing in the home. In both the laboratory and home case studies, sensitivity was poor for few positions for a few participants. Although it was encouraging that those cases did not preclude us from observing inter- and intra-individual differences, more testing—particularly in a larger set of home participants—will be needed to know how robustly our method can deal with poor classifications. Whereas, outliers in many measures used to assess individual differences must simply be trimmed based on a distributional assumption (e.g., extreme CDI scores, Walle and Campos, 2014), our method relies on collecting ground truth data for every individual. Since each individual infant's model can be validated, we have a principled way of excluding outliers based on the prediction accuracy for each infant, each body position, and each session. But, training individual models comes with a cost: It relies on collecting video data for every participant, training those videos, and fitting individual models. It is possible that when a larger set of training data are available, that the accuracy of group models will approach that of individual models. Or, sub-group models could be made to make predictions in infants of the same age (or who share the same repertoire of motor behaviors). Unfortunately, insufficient data were collected from infants of different age groups to test a sub-group approach.

Regardless, future work should investigate why those fits were poor with an eye toward reducing erroneous predictions (instead of excluding data *post hoc*). One possibility is that not enough data were available to train the model for those positions. Although we attempted to elicit different body positions in every infant, infants were not always cooperative. For example, infants who can crawl and walk may be unhappy lying on their backs for minutes at a time. As the time of recording becomes longer, it also creates greater opportunity for errors to arise (such as a caregiver putting on the leggings the wrong way after a diaper change or nap). We hope that by asking caregivers to document such events, we will be able to exclude portions of the day with erroneous data. In the future, collecting validation data (with video) intermittently through the day or at the end of a session could provide a more objective way to check the robustness of the classifier. Given the complexity of testing behavior in the wild, decrements in accuracy for the case study participants (from 98% in the laboratory to 85% in the home) were to be expected. Although it is encouraging that accuracy was still at an acceptable level in the case study participants, more data will be needed to demonstrate whether the method is accurate across a larger sample of participants in the home. Individual differences in infants' motor repertoires and daily routines/activities likely add to heterogeneity in body position frequency, and whether such variability can be captured across a large sample in the home remains to be tested.

Generalizing from training data—a portion of which contained elicited positions—to unconstrained, free-flowing behavior is a significant challenge. As noted, it is especially

difficult when sufficient data for all categories to train and test the models are not available for every infant. One strategy that we used to deal with the unpredictable nature of infant data collection was to design a two-part training procedure—a guided task that attempted to gather data from a fixed list of behaviors followed by a free-play procedure that gathered data from infants in more free-flowing, self-selected positions. Ideally, this two-pronged approach would provide complementary data: In the guided section, the caregiver would place infants in positions that would be rare in free play, such as holding infants and restraining them in a high chair, and free play would capture more naturalistic behavior. However, a limitation of this approach is that we trained and tested models using both guided and free play data. A stronger test would have been to assess model performance on a set of completely naturalistic data (such as a period of free play or home life that excluded any elicited behaviors). Because our approach relied on training models using both types of data, we could not do this in our dataset—there was not enough free play data collected to hold it in reserve for testing. In future work, collecting a separate set of naturalistic testing data would provide a more stringent test of how well models will generalize to body position in daily life.

In addition to providing proof-of-concept data, our two home case studies also highlight the utility of a contactless equipment drop-off procedure for studying infant home behavior. Many infant development researchers—especially those who use looking time metrics—can turn to video conferencing or toolboxes such as Lookit (Scott and Schulz, 2017) for a substitute for in-person studies. In contrast, for researchers who study gross motor behaviors, such as walking and crawling, it may be difficult or impossible to make the paradigm fit on a computer screen. Cameras fixed on a tripod are not ideal for capturing motor behavior, which is why home observation studies typically rely on an experimenter to record infants as they move from place to place (Karasik et al., 2011). Although the 360° cameras we used in the home case studies cannot follow the infant from room to room, they do provide a way to digitally pan and follow the infant. Moreover, the sensors themselves move with infants from place to place, obviating the need for an experimenter to follow infants around. There is no doubt that this method would be easier to implement in person. Although caregivers successfully placed the cameras and leggings on infants, having an experimenter in the home would reduce the burden on the caregiver. In the ideal scenario, the experimenter would briefly visit the home to place the equipment, and then data could be recorded for the rest of the day without the experimenter present.

In summary, characterizing the inputs for development—what infants do and experience on a daily basis—strengthens our ability to build theories (Dahl, 2017; Oakes, 2017; Franchak, 2020). We identified a new way of capturing one type of input, body position, and expect that measuring daily body position experiences will help reveal how infants' burgeoning motor skills are linked with cascading effects on language and spatial cognition (Soska et al., 2010; Oudgenoeg-Paz et al., 2012, 2015; Walle and Campos, 2014; West et al., 2019). In the future, wearable sensors may be used to build machine learning classifiers for other behaviors, such as locomotion (time spent

crawling and walking) and manual activities. In combination with other wearable equipment, such as “headcams” and audio recorders, we may better understand how infants shape the multi-modal inputs for learning through their own actions.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: OSF repository: <https://osf.io/wcga9>, doi: 10.17605/OSF.IO/WCGA9.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board of the University of California, Riverside. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

REFERENCES

- Adolph, K. E., and Franchak, J. M. (2017). The development of motor behavior. *WIREs Cogn. Sci.* 8:e1430. doi: 10.1002/wcs.1430
- Adolph, K. E., and Tamis-LeMonda, C. S. (2014). The costs and benefits of development: The transition from crawling to walking. *Child Dev. Perspect.* 8, 187–192. doi: 10.1111/cdep.12085
- Airaksinen, M., Räsänen, O., Ilén, E., Häyrynen, T., Kivi, A., Marchi, V., et al. (2020). Automatic posture and movement tracking of infants with wearable movement sensors. *Sci. Rep.* 10, 1–13. doi: 10.1038/s41598-019-56862-5
- Arif, M., and Kattan, A. (2015). Physical activities monitoring using wearable acceleration sensors attached to the body. *PLoS ONE* 10:e0130851. doi: 10.1371/journal.pone.0130851
- Armstrong, B., Covington, L. B., Hager, E. R., and Black, M. M. (2019). Objective sleep and physical activity using 24-hour ankle-worn accelerometry among toddlers from low-income families. *Sleep Health* 5, 459–465. doi: 10.1016/j.sleh.2019.04.005
- Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., and Tor, S. (2019). Day by day, hour by hour: naturalistic language input to infants. *Dev. Sci.* 22:e12715. doi: 10.1111/desc.12715
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324
- Bruijns, B. A., Truelove, S., Johnson, A. M., Gilliland, J., and Tucker, P. (2020). Infants' and toddlers' physical activity and sedentary time as measured by accelerometry: a systematic review and meta-analysis. *Int. J. Behav. Nutr. Phys. Act.* 17, 14. doi: 10.1186/s12966-020-0912-4
- Cliff, D. P., Reilly, J. J., and Okely, A. D. (2009). Methodological considerations in using accelerometers to assess habitual physical activity in children aged 0–5 years. *J. Sci. Med. Sport* 12, 557–567. doi: 10.1016/j.jsams.2008.10.008
- Cychosz, M., Romeo, R., Soderstrom, M., Scaff, C., Ganek, H., Cristia, A., et al. (2020). Longform recordings of everyday life: ethics for best practices. *Behav. Res. Methods* 52, 1951–1969. doi: 10.3758/s13428-020-01365-9
- Dahl, A. (2017). Ecological commitments: why developmental science needs naturalistic methods. *Child Dev. Perspect.* 11, 79–84. doi: 10.1111/cdep.12217
- de Barbaro, K. (2019). Automated sensing of daily activity: a new lens into development. *Dev. Psychobiol.* 61, 444–464. doi: 10.1002/dev.21831
- de Barbaro, K., and Fausey, C. M. (2021). Ten lessons about infants' everyday experiences. *PsyArXiv*. doi: 10.31234/osf.io/qa73d
- Fausey, C. M., Jayaraman, S., and Smith, L. B. (2015). “The changing rhythms of life: activity cycles in the first two years of everyday experience,” in *2015 meeting of the Society for Research in Child Development* (Philadelphia, PA).

AUTHOR CONTRIBUTIONS

JF, VS, and CL contributed to the conception and design of the study. VS collected and coded study data. JF completed the statistical analyses. All authors wrote sections of the manuscript, revised, read, and approved the submitted version.

FUNDING

This project was funded by National Science Foundation Grant BCS-1941449 to JF.

ACKNOWLEDGMENTS

The authors thank Brianna McGee and the members of the Perception, Action, and Development Lab for their help in collecting and coding the data. We are grateful to Beth Smith for providing advice about inertial sensors. Finally, we thank the families who participated for making this research possible.

- Franchak, J. M. (2019). Changing opportunities for learning in everyday life: infant body position over the first year. *Infancy* 24, 187–209. doi: 10.1111/inf.12272
- Franchak, J. M. (2020). The ecology of infants' perceptual-motor exploration. *Curr. Opin. Psychol.* 32, 110–114. doi: 10.1016/j.copsyc.2019.06.035
- Franchak, J. M., Kretch, K. S., and Adolph, K. E. (2018). See and be seen: infant-caregiver social looking during locomotor free play. *Dev. Sci.* 21:e12626. doi: 10.1111/desc.12626
- Franchak, J. M., Kretch, K. S., Soska, K. C., and Adolph, K. E. (2011). Head-mounted eye tracking: a new method to describe infant looking. *Child Dev.* 82, 1738–1750. doi: 10.1111/j.1467-8624.2011.01670.x
- Gibson, E. J. (1988). Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. *Annu. Rev. Psychol.* 39, 1–41. doi: 10.1146/annurev.ps.39.020188.000245
- Greenspan, B., Cunha, A. B., and Lobo, M. A. (2021). Design and validation of a smart garment to measure positioning practices of parents with young infants. *Infant. Behav. Dev.* 62:101530. doi: 10.1016/j.infbeh.2021.101530
- Hagenbuchner, M., Cliff, D. P., Trost, S. G., Van Tuc, N., and Peoples, G. E. (2015). Prediction of activity type in preschool children using machine learning techniques. *J. Sci. Med. Sport* 18, 426–431. doi: 10.1016/j.jsams.2014.06.003
- Hager, E., Tilton, N., Wang, Y., Kapur, N., Arbaiza, R., Merry, B., et al. (2017). The home environment and toddler physical activity: an ecological momentary assessment study. *Pediatr. Obes.* 12, 1–9. doi: 10.1111/ijpo.12098
- He, M., Walle, E. A., and Campos, J. J. (2015). A cross-national investigation of the relationship between infant walking and language development. *Infancy* 20, 283–305. doi: 10.1111/inf.12071
- Hewitt, L., Stanley, R. M., Cliff, D., and Okely, A. D. (2019). Objective measurement of tummy time in infants (0–6 months): a validation study. *PLoS ONE* 14:e0210977. doi: 10.1371/journal.pone.0210977
- Hnatiuk, J., Salmon, J., Campbell, K. J., Ridgers, N. D., and Hesketh, K. D. (2013). Early childhood predictors of toddlers' physical activity: longitudinal findings from the Melbourne infant program. *Int. J. Behav. Nutr. Phys. Act.* 10:e123. doi: 10.1186/1479-5868-10-123
- Jiang, C., Lane, C. J., Perkins, E., Schiesel, D., and Smith, B. A. (2018). Determining if wearable sensors affect infant leg movement frequency. *Dev. Neurorehabil.* 21, 133–136. doi: 10.1080/17518423.2017.1331471
- Kadooka, K., Caufield, M., Fausey, C. M., and Franchak, J. M. (2021). “Visuomotor learning opportunities are nested within everyday activities,” in *Paper Presented at the Biennial Meeting of the Society for Research in Child Development*.
- Karasik, L. B., Tamis-LeMonda, C. S., and Adolph, K. E. (2011). Transition from crawling to walking and infants' actions with objects and people. *Child Dev.* 82, 1199–1209. doi: 10.1111/j.1467-8624.2011.01595.x

- Karasik, L. B., Tamis-LeMonda, C. S., and Adolph, K. E. (2014). Crawling and walking infants elicit different verbal responses from mothers. *Dev. Sci.* 17, 388–395. doi: 10.1111/desc.12129
- Karasik, L. B., Tamis-LeMonda, C. S., Adolph, K. E., and Bornstein, M. H. (2015). Places and postures: a cross-cultural comparison of sitting in 5-month-olds. *J. Cross Cult. Psychol.* 46, 1023–1038. doi: 10.1177/0022022115593803
- Kretch, K. S., Franchak, J. M., and Adolph, K. E. (2014). Crawling and walking infants see the world differently. *Child Dev.* 85, 1503–1518. doi: 10.1111/cdev.12206
- Kuzik, N., Clark, D., Ogden, N., Harber, V., and Carson, V. (2015). Physical activity and sedentary behaviour of toddlers and preschoolers in child care centres in Alberta, Canada. *Can. J. Public Health* 106, e178–e183. doi: 10.17269/cjph.106.4794
- Kwon, S., Zavos, P., Nickelle, K., Sugianto, A., and Albert, M. V. (2019). Hip and wrist-worn accelerometer data analysis for toddler activities. *Int. J. Environ. Res. Public Health* 16, 2598. doi: 10.3390/ijerph16142598
- Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310
- Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R News* 2, 18–22.
- Libertus, K., and Hauf, P. (2017). Motor skills and their foundational role for perceptual, social, and cognitive development. *Front. Psychol.* 8:301. doi: 10.3389/fpsyg.2017.00301
- Lobo, M. A., Hall, M. L., Greenspan, B., Rohloff, P., Prosser, L. A., and Smith, B. A. (2019). Wearables for pediatric rehabilitation: how to optimally design and use products to meet the needs of users. *Phys. Ther.* 99, 647–657. doi: 10.1093/ptj/pzz024
- Luo, C., and Franchak, J. M. (2020). Head and body structure infants' visual experiences during mobile, naturalistic play. *PLoS ONE* 15:e0242009. doi: 10.1371/journal.pone.0242009
- Majnemer, A., and Barr, R. G. (2005). Influence of supine sleep positioning on early motor milestone acquisition. *Dev. Med. Child Neurol.* 47, 370–376. doi: 10.1017/S0012162205000733
- Moore, C., Dailey, S., Garrison, H., Amatuni, A., and Bergelson, E. (2019). Point, walk, talk: links between three early milestones, from observation and parental report. *Dev. Psychol.* 55, 1579–1593. doi: 10.1037/dev0000738
- Nam, Y., and Park, J. W. (2013). Child activity recognition based on cooperative fusion model of a triaxial accelerometer and a barometric pressure sensor. *IEEE J. Biomed. Health Inform.* 17, 420–426. doi: 10.1109/JBHI.2012.2235075
- Nickel, L. R., Thatcher, A. R., Keller, F., Wozniak, R. H., and Iverson, J. M. (2013). Posture development in infants at heightened versus low risk for autism spectrum disorders. *Infancy* 18, 639–661. doi: 10.1111/inf.12025
- Oakes, L. M. (2017). Plasticity may change inputs as well as processes, structures, and responses. *Cogn. Dev.* 42:4–14. doi: 10.1016/j.cogdev.2017.02.012
- Oudgenoeg-Paz, O., Leseman, P. P. M., and Volman, M. C. J. M. (2015). Exploration as a mediator of the relation between the attainment of motor milestones and the development of spatial cognition and spatial language. *Dev. Psychol.* 51, 1241–1253. doi: 10.1037/a0039572
- Oudgenoeg-Paz, O., Volman, M. C. J. M., and Leseman, P. P. M. (2012). Attainment of sitting and walking predicts development of productive vocabulary between ages 16 and 28 months. *Infant Behav. Dev.* 35, 733–736. doi: 10.1016/j.infbeh.2012.07.010
- Preece, S. J., Goulermas, J. Y., Kenney, L. P. J., and Howard, D. (2009). A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. *IEEE Trans. Biomed. Eng.* 56, 871–879. doi: 10.1109/TBME.2008.2006190
- Ren, X., Ding, W., Crouter, S. E., Mu, Y., and Xie, R. (2016). Activity recognition and intensity estimation in youth from accelerometer data aided by machine learning. *Appl. Intell.* 45, 512–529. doi: 10.1007/s10489-016-0773-3
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Smith, B. A., Vanderbilt, D. L., Applequist, B., and Kyvelidou, A. (2017). Sample entropy identifies differences in spontaneous leg movement behavior between infants with typical development and infants at risk of developmental delay. *Technologies* 5, 55. doi: 10.3390/technologies5030055
- Soska, K. C., and Adolph, K. E. (2014). Postural position constrains multimodal object exploration in infants. *Infancy* 19, 138–161. doi: 10.1111/inf.12039
- Soska, K. C., Adolph, K. E., and Johnson, S. P. (2010). Systems in development: Motor skill acquisition facilitates three-dimensional object completion. *Dev. Psychol.* 46, 129–138. doi: 10.1037/a0014618
- Stewart, T., Narayanan, A., Hedayatrad, L., Neville, J., Mackay, L., and Duncan, S. (2018). A dual-accelerometer system for classifying physical activity in children and adults. *Med. Sci. Sports Exerc.* 50, 2595–2602. doi: 10.1249/MSS.0000000000001717
- Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* 14, 323–348. doi: 10.1037/a0016973
- Tamis-LeMonda, C. S., Kuchirko, Y., Luo, R., Escobar, K., and Bornstein, M. H. (2017). Power in methods: language to infants in structured and naturalistic contexts. *Dev. Sci.* 20:e12456. doi: 10.1111/desc.12456
- Thurman, S. L., and Corbetta, D. (2017). Spatial exploration and changes in infant-mother dyads around transitions in infant locomotion. *Dev. Psychol.* 53, 1207–1221. doi: 10.1037/dev0000328
- Trost, S., Cliff, D., Ahmadi, M. N., Van Tuc, N., and Hagenbuchner, M. (2018). Sensor-enabled activity class recognition in preschoolers: hip versus wrist data. *Med. Sci. Sports Exerc.* 50, 634–641. doi: 10.1249/MSS.0000000000001460
- Trost, S. G., Fees, B. S., Haar, S. J., Murray, A. D., and Crowe, L. K. (2012). Identification and validity of accelerometer cut-points for toddlers. *Obesity* 20, 2317–2319. doi: 10.1038/oby.2011.364
- Walle, E. A. (2016). Infant social development across the transition from crawling to walking. *Front. Psychol.* 7:e960. doi: 10.3389/fpsyg.2016.00960
- Walle, E. A., and Campos, J. J. (2014). Infant language development is related to the acquisition of walking. *Dev. Psychol.* 50:336–348. doi: 10.1037/a0033238
- Weisleder, A., and Fernald, A. (2013). Talking to children matters: early language experience strengthens processing and builds vocabulary. *Psychol. Sci.* 24, 2143–2152. doi: 10.1177/0956797613488145
- West, K. L., Leezenbaum, N. B., Northrup, J. B., and Iverson, J. M. (2019). The relation between walking and language in infant siblings of children with autism spectrum disorder. *Child Dev.* 90, e356–e372. doi: 10.1111/cdev.12980
- Yao, X., Plötz, T., Johnson, M., and de Barbaro, K. (2019). Automated detection of infant holding using wearable sensing: implications for developmental science and intervention. *Proc. ACM Inter. Mobile Wearable Ubiquitous Technol.* 3, 1–17. doi: 10.1145/3328935
- Zhao, W., Adolph, A. L., Puyau, M. R., Vohra, F. A., Butte, N. F., and Zakeri, I. F. (2013). Support vector machines classifiers of physical activities in preschoolers. *Physiol. Rep.* 1, e00006. doi: 10.1002/phy2.6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Franchak, Scott and Luo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Comparing Face-to-Face and Online Data Collection Methods in Preterm and Full-Term Children: An Exploratory Study

Paige M. Nelson^{1*}, Francesca Scheiber¹, Haley M. Laughlin¹ and Ö. Ece Demir-Lira^{1,2,3}

¹ Department of Psychological and Brain Sciences, The University of Iowa, Iowa City, IA, United States, ² Delta Center, The University of Iowa, Iowa City, IA, United States, ³ Iowa Neuroscience Institute, The University of Iowa, Iowa City, IA, United States

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Rachel M. Flynn,
San Francisco State University,
United States
Cristina de-la-Peña,
Universidad Internacional de La Rioja,
Spain

*Correspondence:

Paige M. Nelson
paige-nelson@uiowa.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 June 2021

Accepted: 08 October 2021

Published: 28 October 2021

Citation:

Nelson PM, Scheiber F,
Laughlin HM and Demir-Lira ÖE
(2021) Comparing Face-to-Face
and Online Data Collection Methods
in Preterm and Full-Term Children: An
Exploratory Study.
Front. Psychol. 12:733192.
doi: 10.3389/fpsyg.2021.733192

The COVID-19 pandemic has transformed the landscape for children's daily lives and the landscape for developmental psychology research. Pandemic-related restrictions have also significantly disrupted the traditional face-to-face methods with which developmental scientists produce research. Over the past year, developmental scientists have published on the best practices for online data collection methods; however, existing studies do not provide empirical evidence comparing online methods to face-to-face methods. In this study, we tested feasibility of online methods by examining performance on a battery of standardized and experimental cognitive assessments in a combined sample of 4- to 5-year-old preterm and full-term children, some of whom completed the battery face-to-face, and some of whom completed the battery online. First, we asked how children's performance differs between face-to-face and online format on tasks related to verbal comprehension, fluid reasoning, visual spatial, working memory, attention and executive functioning, social perception, and numerical skills. Out of eight tasks, we did not find reliable differences on five of them. Second, we explored the role of parent involvement in children's performance in the online format. We did not find a significant effect of parent involvement on children's performance. Exploratory analyses showed that the role of format did not vary for children at risk, specifically children born preterm. Our findings contribute to the growing body of literature examining differences and similarities across various data collection methods, as well as literature surrounding online data collection for continuing developmental psychology research.

Keywords: COVID-19, in-person data collection, online data collection, children, prematurity, neurocognitive assessment, developmental psychology

INTRODUCTION

The COVID-19 pandemic has transformed the landscape for children's daily lives and the landscape for developmental psychology research. Schools across the world have restructured and developed online learning curriculums. Around 214 million children are estimated to have missed more than three quarters of in-person education in 2020 (United Nations Children's Fund, 2021). More than 90% of children in the United States are estimated to have received some form of distance learning

during COVID-19 (Bureau, 2021). Likewise, the provision of community and social services has suffered, despite being needed now more than ever (Tsega et al., 2020; EducationData, 2021). Pandemic-related restrictions have also significantly disrupted the traditional methods in which developmental scientists produce research—that is, in-person studies requiring face-to-face interactions (Garrisi et al., 2020). Given the likely continued role of online assessments in research and clinical services, it is important to understand both the differences and similarities between children's performance in face-to-face and online settings. The direct and long-term effects of online measurement of children's performance remain largely unknown. The goal of the current paper is to add to this growing body of literature by testing the feasibility of online data collection methods and comparing 4- to 5-year-old preschoolers' performance face-to-face vs. online on a wide variety of standardized and experimental cognitive assessments.

Over the past few years, developmental scientists have published on the best practices for online data collection methods (Frank et al., 2016; Garrisi et al., 2020; Lourenco and Tasimi, 2020; Manning et al., 2020; Nussenbaum et al., 2020; Rhodes et al., 2020; Sheskin et al., 2020; Morini and Blair, 2021; Su and Ceci, 2021). These studies agree that while online data collection methods are still in their infancy, online measurements have become a promising platform for developmental psychology research. However, only a few studies provide empirical evidence comparing online methods to face-to-face methods. For example, Morini and Blair (2021) examined the feasibility of collecting remote eye-gaze data with children. They compared their online sample to a previously collected face-to-face sample, during which they found their online data collection methods to be reliable and sufficient in conducting developmental language research (Morini and Blair, 2021).

An emerging body of work also focuses on the reliability and validity of online data collection methods. For example, Manning et al. (2020) examined the feasibility, reliability, and validity of child language samples drawn from recorded parent-child interactions via video chat. They found child language samples (i.e., key speech and language measures) collected via video chat vs. face-to-face laboratory video recordings to be comparable (Manning et al., 2020). So far, studies have focused on narrow aspects of children's performance, thus it is important to examine children's performance on a wide array of standardized and experimental measures assessing multiple cognitive domains to gain a more complete view of face-to-face vs. online assessments. Our primary goal is to examine how children's performance in verbal comprehension, fluid reasoning, visual spatial, working memory, attention and executive functioning, social perception, and numerical tasks differ between a lab based, face-to-face format, and an online format.

According to some, remote research has many benefits, including its ability to broaden sample diversity, as compared to face-to-face laboratory studies (Lourenco and Tasimi, 2020). For example, online methods might be more inclusive of groups who do not or cannot attend face-to-face research studies, including atypically developing children. Given the greater need for assessments and interventions for at-risk children in clinical settings, it is fundamental to better understand whether

and how online interactions influence children with atypical developmental trajectories. Children born preterm (<37 weeks gestational age) fall into such an "at risk" group. Every year, close to 15 million children in the United States are born preterm (Wolke et al., 2019). Those who survive have an increased risk for death, disability, and delay (Centers for Disease Control and Prevention [CDC], 2020). Preterm-born children (PTB) fall behind term-born children (TB) on various measures of cognitive performance (Allotey et al., 2018; Brydges et al., 2018), with children born the earliest tending to have the worst outcomes (Bhutta et al., 2002; Snijders et al., 2020). Likewise, this gap in cognitive performance between PTB and TB children often persists throughout formal schooling. Better understanding how at-risk children perform in face-to-face vs. online formats will have implications for future assessment and intervention efforts. Thus, in the current study, we diversify our sample by including both PTB children and TB children.

Taken together, our goal is to contribute to the growing literature on establishing the reliability of online research methods by examining children's performance on standardized and experimental cognitive assessments. We examine performance in a combined sample of 4- to 5-year-old TB and PTB children, some of whom completed the battery face-to-face and some of whom completed the battery online. We ask how children's performance differs between face-to-face and online format on tasks related to verbal comprehension, fluid reasoning, visual spatial, working memory, attention and executive functioning, social perception, and numerical skills. We supplement our main research question with two exploratory analyses. Online data collections methods typically rely on parents, but how parent involvement during remote data collection influences children's performance has yet to be explored. To address this question, we explore the role of parent involvement in children's performance in the online format. How children's performance differs between face-to-face and online format on tasks related to verbal comprehension, fluid reasoning, visual spatial, working memory, attention and executive functioning, social perception, and numerical skills as a function of a prematurity also has yet to be explored. To address this question as an exploratory aim, we examine the role of prematurity in children's performance in both the face-to-face and online format.

METHODS

Participants

Participants were 93 TB (≥ 37 weeks gestational age) and 38 PTB (<37 weeks gestational age), for a total of 131 children, who participated in an ongoing longitudinal study on the relations between preterm birth and neurodevelopment. Fifty-four TB children and 29 PTB completed the study face-to-face in a lab-based format. Thirty-nine TB children and nine PTB completed the study in an online format via Zoom video conferencing. Overall, 83 children (29 PTB) completed the study face-to-face in a lab-based format and 48 children (9 PTB) completed the study in an online format. This study was approved by the Institutional Review Board at our local university. We recruited parent-child

dyads using the university hospital's electronic health records, university mass emailing, social media, and word of mouth. Parent-child dyads were eligible for this study if the child was between the ages of 4 and 5 years old, was a native speaker of English, had normal or corrected-to-normal vision and hearing, had no history of a genetic syndrome or birth defect, and had no limitations (based on parental report) that would prevent them from completing paper/pencil tasks. For those who completed the study in an online format, it was also preferred that they had an electronic device (computer, laptop, tablet, or smart phone) with reliable internet. Parent-child dyads without an electronic device were mailed an Amazon Fire tablet that they could use to participate in the online sessions. We began enrollment for the face-to-face study in June 2019 and paused data collection in March 2020, due to the COVID-19 pandemic. In October 2020, we began the enrollment for the online study, for which we used Zoom video conferencing. The online data collection is ongoing; for the purposes of the current manuscript, we report on data collected through mid-June 2021.

Table 1 shows demographic characteristics for the face-to-face and online samples. Face-to-face and online parent-child dyads did not significantly differ in child age, gender, ethnicity, race, gestational age, birthweight, parent education, or household income. Children and parents were predominately White and from high-socioeconomic backgrounds, with an average household income of \$115,648.71 and an average parent education corresponding to a college degree.

Procedure

For the face-to-face portion of the study, parent-child dyads attended a 3-h laboratory visit. During the laboratory visit, experimenters administered tasks to children, while parents

completed questionnaires on a computer in another room. The face-to-face portion of the study included ten standardized neurocognitive assessments and four experimental tasks. The standardized neurocognitive assessments included six subtests from the Wechsler Preschool & Primary Scale of Intelligence, Fourth Edition (WPPSI-IV; Wechsler, 2012) (block design, bug search, matrix reasoning, information, similarities, and picture memory) and four subtests from the Developmental Neuropsychological Assessment, Second Edition (NEPSY-II; Brooks et al., 2009) (affect recognition, comprehension of instruction, statue, and theory of mind). The three experimental tasks included Give A Number, What's on This Card, and Mental Rotation. Tasks were administered in blocks, and children took breaks in between each block.

For the online portion of the study, parent-child dyads participated in four 45–60-min sessions via Zoom. Using feedback from a focus group with local parents who expressed concern regarding possible screen fatigue, we structured the online portion of the study across four, shorter online sessions rather than one 3-h session. During the online sessions, children completed most the same standardized neurocognitive assessments, and parents completed the same online questionnaires. In the online portion of the study, parents were asked to complete the same questionnaires on their own time between session 1 and session 4. Four of the neurocognitive assessments (WPPSI-IV block design, give a number, NEPSY-II comprehension of instructions, and NEPSY-II theory of mind) that were administered face-to-face could not reliably be administered in an online format for reasons discussed below. The six remaining standardized neurocognitive assessments and two remaining experimental tasks were divided across four online Zoom sessions. Session 1 included WPPSI-IV matrix reasoning, information, and similarities. Session 2 included what's on this card and mental rotation. Session 3 included NEPSY-II affect recognition and statue. Session 4 included WPPSI-IV picture memory. The order of the tasks was the same in both the face-to-face and online sessions, and tasks were administered by the same research assistants.

For the online portion of the study, parents scheduled their four sessions via Calendly (an online scheduling tool), email, or phone. Once parents scheduled their sessions, experimenters provided parent-child dyads with information to prepare them for their first session. This included information on preparing devices and information on dos and don'ts for the four sessions. For example, experimenters stressed the importance of not providing aid or input during the neurocognitive assessments. This also included a link to the informed consent document if the parent did not fill it out prior. The day of session 1, experimenters emailed parents a secured Zoom video conferencing link, for which they were able to attend without needing a Zoom account. Once parent-child dyads logged onto the session, experimenters guided parents on positioning the camera if necessary. With parental consent, all online sessions were recorded through Zoom. The same procedures were followed for session 2 through session 4. Further information on task set up is in **Table 2**.

TABLE 1 | Demographic information for face-to-face ($N = 83$) and online ($N = 48$) samples.

| | Face-to-face <i>M (SD) or n (%)</i> | Online <i>M (SD) or n (%)</i> |
|-------------------------|--|----------------------------------|
| Child age (years) | 4.77 (0.48) | 5.15 (0.48) |
| Child gender | | |
| Female | 38 (46%) | 25 (53%) |
| Child hispanic | 7 (8%) | 4 (9%) |
| Child white | 80 (96%) | 45 (94%) |
| Child premature | 29 (35%) | 9 (19%) |
| Birth weight (lbs, oz) | 6.20 (2.42) | 6.81 (1.71) |
| Household income (USD) | 113233 (67708) | 120204 (56229) |
| Parent education | | |
| High school graduate | 3 (4%) | 2 (4%) |
| Some college credit | 5 (6%) | 6 (13%) |
| Associate's degree | 10 (12%) | 4 (9%) |
| Bachelor's degree | 37 (45%) | 10 (21%) |
| Professional degree | 28 (34%) | 25 (53%) |
| Parent age (years) | 36.59 (4.87) | 38.61 (13.04) |
| Parent gender | | |
| Female | 75 (91%) | 43 (93%) |
| Parent hispanic | 3 (4%) | 1 (2%) |
| Parent white | 77 (93%) | 45 (94%) |

TABLE 2 | Task descriptions for face-to-face and online procedures.

| Task | Included face-to-face? | Included online? | Timing/response type for online | Face-to-face procedure | Online procedure |
|--|------------------------|------------------|---------------------------------|--|--|
| Verbal comprehension | | | | | |
| WPPSI-IV information | ✓ | ✓ | Untimed / verbal response | For questions involving pictures, experimenters presented children with visual stimuli via a testing binder, and flipped the pages when children were ready to move items. Children provided their response by pointing. For questions involving verbal response, experimenters read questions aloud, and children responded verbally. | For questions involving pictures, experimenters presented children with visual stimuli via Microsoft PowerPoint; experimenters shared screen and advanced slides when children were ready to move items. Children provided their response by 'stamping' via Zoom's remote-control feature, on the experimenter's screen using a computer mouse or trackpad. Parents were asked to assist, during which they helped their child 'stamp' their response. For questions involving verbal response, participants were administered the Information subtest online using the same method as in-person. |
| WPPSI-IV similarities | ✓ | ✓ | Untimed / verbal response | For questions involving pictures, experimenters presented children with visual stimuli via a testing binder, and flipped the pages when children were ready to move items. Children provided their response by pointing. For questions involving verbal response, experimenters read questions aloud, and children responded verbally. | For questions involving pictures, experimenters presented children with visual stimuli via Microsoft PowerPoint; experimenters shared screen and advanced slides when children were ready to move items. Children provided their response by 'stamping' via Zoom's remote-control feature, on the experimenter's screen using a computer mouse or trackpad. Parents were asked to assist, during which they helped their child 'stamp' their response. For questions involving verbal response, participants were administered the Similarities subtest online using the same method as in-person. |
| Language | | | | | |
| NEPSY-II comprehension of instructions | ✓ | | Untimed / response stamp | Experimenters presented children with visual stimuli via a testing binder, during which experimenters gave verbal instructions that increased in complexity and could not be repeated. Children provided their response by pointing. | |
| Visuospatial | | | | | |
| WPPSI-IV block design | ✓ | | | Experimenters presented children with blocks of various colors and patterns and asked children to model 3-D block patterns of increased complexity | |
| Mental rotation | ✓ | ✓ | Untimed / response stamp | Experimenters presented children with visual stimuli via a testing binder and flipped the pages when children were ready to move items. Children provided their response by pointing. | Experimenters presented children with visual stimuli via Microsoft PowerPoint; experimenters shared screen and advanced slides when children were ready to move items. Children provided their response by 'stamping' via Zoom's remote-control feature, on the experimenter's screen using a computer mouse or trackpad. Parents were asked to assist, during which they helped their child 'stamp' their response. |
| Fluid reasoning | | | | | |
| WPPSI-IV matrix reasoning | ✓ | ✓ | Untimed / response stamp | Experimenters presented children with visual stimuli via a testing binder and flipped the pages when children were ready to move items. Children provided their response by pointing. | Experimenters presented children with visual stimuli via Microsoft PowerPoint; experimenters shared screen and advanced slides when children were ready to move items. Children provided their response by 'stamping' via Zoom's remote-control feature, on the experimenter's screen using a computer mouse or trackpad. Parents were asked to assist, during which they helped their child 'stamp' their response. |

(Continued)

TABLE 2 | (Continued)

| Task | Included face-to-face? | Included online? | Timing/response type for online | Face-to-face procedure | Online procedure |
|--|------------------------|------------------|--|--|--|
| Working memory | | | | | |
| WPPSI-IV picture memory | ✓ | ✓ | Timed / response stamp | Experimenters presented children with visual stimuli via a testing binder and flipped the pages when children were ready to move items. Children provided their response by pointing. | Experimenters presented children with visual stimuli via Microsoft PowerPoint; experimenters shared screen and advanced slides when children were ready to move items. Children provided their response by 'stamping' via Zoom's remote-control feature, on the experimenter's screen using a computer mouse or trackpad. Parents were asked to assist, during which they helped their child 'stamp' their response. |
| Attention and executive functioning | | | | | |
| NEPSY-II statue | ✓ | ✓ | Timed | Experimenters administered the Statue subtest online using the same method as in-person, with the exception that experimenters often asked parents to help orient their child in the position of the camera. Otherwise, parents were not required to assist on the Statue subtest. | |
| Social perception | | | | | |
| NEPSY-II affect recognition | ✓ | ✓ | Untimed / response stamp | Experimenters presented children with visual stimuli via a testing binder and flipped the pages when children were ready to move items. Children provided their response by pointing. | Experimenters presented children with visual stimuli via Microsoft PowerPoint; experimenters shared screen and advanced slides when children were ready to move items. Children provided their response by 'stamping' via Zoom's remote-control feature, on the experimenter's screen using a computer mouse or trackpad. Parents were asked to assist, during which they helped their child 'stamp' their response. |
| NEPSY-II theory of mind | ✓ | | Untimed / verbal response / response stamp | For the Verbal task and contextual tasks, experimenters presented children with visual stimuli via a testing binder and flipped the pages when children were ready to move items. For the Contextual task, children provided their response by pointing. | |
| Numerical | | | | | |
| What's on this card | ✓ | ✓ | Untimed / verbal response | Experimenters presented children with visual stimuli via a testing binder and flipped the pages when children were ready to move items. Children provided their response by responding verbally. | Experimenters presented children with visual stimuli via Microsoft PowerPoint; experimenters shared screen and advanced slides when children were ready to move items. Children provided their response by responding verbally. Parents were not required to assist on the WOC task. |
| Give a number | ✓ | | | Experimenters presented children with a pile of fifteen plastic fish, during which experimenters asked children to place a certain number of fish (i.e., 1–9) into a fishbowl. | |
| Processing speed | | | | | |
| WPPSI-IV bug search | ✓ | | Timed / response stamp | Experimenters presented children with visual stimuli via a testing packet and flipped the pages when children were ready to move items. Experimenters asked the children to match various kinds of bugs to one another from an assortment of response options using a child-friendly ink dauber, within one minute and 15 seconds. | |

Measures Administered Both Face-to-Face and Online

Experimenters administered the following measures. It should be noted that publishers of standardized neurocognitive assessments

did not provide the online materials; we adapted the materials for online administration for research purposes. This is true for all WPPSI-IV and NEPSY-II materials. The WPPSI-IV and NEPSY-II have been shown to have strong reliability and validity

when measured in a face-to-face format (Brooks et al., 2009; Wechsler, 2012). We also measured internal consistency using Cronbach's alpha in our two experimental tasks. Reliability across face-to-face and online participants on What's on This Card was good ($\alpha > 0.75$). Reliability across face-to-face participants on Mental Rotation was acceptable ($\alpha = 0.50$), and across online participants, reliability was good ($\alpha = 0.74$). Please see **Table 2** for further details on the tasks referenced below.

Verbal Comprehension

WPPSI-IV information

Experimenters administered the Information subtest, part of the WPPSI-IV Verbal Comprehension Index (VCI). The VCI measures a child's acquired knowledge, verbal reasoning, and comprehension skills. The Information subtest uses both visual and verbal stimuli to assess children's acquired knowledge (e.g., "what do people use to stay dry in the rain?") (Wechsler, 2012).

WPPSI-IV similarities

Experimenters administered the Similarities subtest, part of the WPPSI-IV VCI. The Similarities subtest asks children, using Picture Tasks and Verbal Tasks, to describe how two words that share a common characteristic are related to one another (e.g., "red and yellow are both. . .") (Wechsler, 2012).

Fluid Reasoning

WPPSI-IV matrix reasoning

Experimenters administered the Matrix Reasoning (WPPSI-MR) subtest, part of the WPPSI-IV Fluid Reasoning Index (FRI). WPPSI-MR measures visual processing and spatial perception by asking children to select a missing portion from a matrix (Wechsler, 2012).

Visual Spatial

Mental rotation

Experimenters administered a shortened version of the Children's Mental Transformation Task (CMTT, Levine et al., 1999). The Children's Mental Transformation Task is a non-verbal spatial task, during which children are presented with four shapes and two halves of a 2D shape and asked to select the shape that the two halves would make if they were put together.

Working Memory

WPPSI-IV picture memory

Experimenters administered the Picture Memory (WPPSI-PM) subtest, part of the WPPSI-IV Working Memory Index (WMI). WPPSI-PM measures a child's working memory by asking children to look at pictures of increasingly complex quantities, for three seconds, before asking children to point to those they viewed on a response page (Wechsler, 2012).

Attention and Executive Functioning

NEPSY-II statue

Experimenters administered the Statue subtest, part of the NEPSY-II Attention and Executive Functioning domain, which measures a child's motor persistence and inhibition (Brooks et al., 2009). Experimenters asked children to remain as still as possible,

during which experimenters deducted points if children opened their eyes, made drastic body movements, and/or spoke.

Social Perception

NEPSY-II affect recognition

Experimenters administered the affect recognition (AR) subtest, part of the NEPSY-II Social Perception domain. The AR subtest measures a child's ability to recognize affect (Brooks et al., 2009). In four different tasks, experimenters showed children variations of affect from photographs of children's faces, during which experimenters assessed children's ability to recognize affect between children in each task (Brooks et al., 2009).

Numerical

What's on this card

Experimenters administered a What's on This Card task, during which experimenters asked children to vocalize what they saw on twelve different cards. For example, experimenters showed a card with three soccer balls and asked children "what is on this card?" Experimenters scored the total amount of cards the participant responded correctly to, out of a total of seventeen.

Measures Administered Face-To-Face but Not Online

We were unable to move several face-to-face measures to an online format. Some tasks were excluded because we were not able to provide the necessary materials to each participant. For example, we could not administer the WPPSI-IV block design (BD) subtest, part of the WPPSI-IV visual spatial index (VSI), in an online format because we were unable to supply the standardized materials (i.e., blocks, assessment binder, and stopwatch) to every participant. Likewise, we could not administer WPPSI-IV bug search (BS) subtest, part of the WPPSI-IV processing speed index (PSI), and an experimental numerical task, give a number task, due to the same reason.

Other tasks were excluded because we were concerned about administering these tasks in a standardized fashion across various electronic devices (i.e., laptop vs. tablet) and internet reliabilities. For example, we chose to not administer the NEPSY-II comprehension of instructions (CI) subtest, part of the NEPSY-II Language domain, due to being unable to repeat the verbal instructions to the child in case a problem with internet connection occurred. We chose to not administer the NEPSY-II theory of mind (TM) subtest, part of the NEPSY-II Social Perception domain, due to the same reason.

Parental Involvement Coding

For online sessions only, we categorized parental involvement into two categories: (1) parent absent or quiet and (2) parent present but with minimal involvement, including directing attention or rewording the instructions. If the parent was present with significant, more than minimal, involvement, the child's score from that task was excluded from the analyses below. Examples of significant involvement included parents providing strategies relevant to the task that would significantly influence child performance or parents simply providing the answer. This

coding was completed after the sessions were done using the video recordings of the interaction.

Analytic Plan

We examined whether children's performance varied as a function of format. A sequence of multiple linear regressions were run using the *lm* function in the *stats* package (R Core Team, 2013) to determine whether children in the online condition performed differently than children administered the battery face-to-face. Format was dummy coded with face-to-face condition used as the reference group. In predicting performance on WPPSI-IV and NEPSY-II, we controlled for parent education only because we used participant's scaled scores adjusted for age at testing. In predicting performance on What's On This Card and Mental Rotation, we controlled for parent education and child age at testing. For exploratory purposes, we also examined if child performance differed in the face-to-face and online format as a function of prematurity (TB vs. PTB) by later adding the interaction component between prematurity and format to the sequence of multiple linear regressions used to answer research question 1. Finally, again for exploratory analyses, we examined if child performance at a given session varied as a function of parental involvement in the session during which task was administered using *t*-tests with involvement as two categories (no involvement vs. minimal involvement).

RESULTS

How Do Children's Performance in Verbal Comprehension, Fluid Reasoning, Visual Spatial, Working Memory, Attention and Executive Functioning, Social Perception, and Numerical Skills Differ From Face-to-Face to Online Format?

Figure 1 and **Table 3** represent children's performance on several measures, both in the face-to-face and in the online study. Face-to-face and online samples differed significantly from each other on the following tasks: WPPSI-IV information, WPPSI-IV similarities, and WPPSI-IV MR. Scores on all three subtests were reliably lower in the face-to-face sample than in the online sample, when controlling for parent education.

Scores on the following standardized neurocognitive assessments did not differ between children who participated face-to-face and children who participated online, when controlling for parent education: WPPSI-IV PM, NEPSY-II AR, and NEPSY-II statue. Scores on the following experimental tasks did not differ between children who participated face-to-face and children who participated online, when controlling for parent education and age at testing: mental rotation and what's on this card.

It could be the case that the lowest performing children were unable to perform the tasks online, resulting in experimenters dropping their scores. To test this possibility, we conducted a follow-up analysis, in which we compared whether the number of children whose data were dropped in the face-to-face vs.

online study differed from each other. Data were excluded for the following reasons: significant parent involvement, child fatigue, and experimenter error. The number of children whose data were included vs. excluded from the analysis are reported in **Table 3** and were compared across the two formats using Chi-Square analyses. None of the comparisons reached statistical significance (all *p*'s > 0.05), suggesting that number of children whose data were excluded due to reasons stated above did not vary across formats.

What Is the Role of Parental Involvement in Children's Performance in the Online Format?

Finally, for the children who participated in the online format only, we examined if parent involvement played a role. To reduce the number of analyses we ran, we only conducted these exploratory analyses on the three WPPSI-IV subtests on which children performed better in the online format as compared to face-to-face format. For children whose data were included in the first session, twenty-five parents were not involved and 22 were minimally involved. *t*-test analyses comparing children of parents who were not involved vs. those who were minimally involved did not reveal any significant differences on WPPSI-IV similarities, $t(42) = 0.08$, $p = 0.94$, WPPSI-IV information, $t(43) = 2.02$, $p = 0.36$, or on WPPSI-IV MR, $t(39) = 1.23$, $p = 0.23$.

How Do Children's Performance in Verbal Comprehension, Fluid Reasoning, Visual Spatial, Working Memory, Attention and Executive Functioning, Social Perception, and Numerical Skills Differ From Face-to-Face to Online Format as a Function of Prematurity?

Scores on all standardized neurocognitive assessments did not differ between children who participated face-to-face and children who participated online as a function of prematurity, except on WPPSI-IV MR. A reliable bivariate interaction emerged between format and prematurity on WPPSI-IV MR, $t(114) = -2.22$, $p < 0.05$. PTB children performed lower than TB children in the online format, as compared to children in the face-to-face format.

DISCUSSION

The COVID-19 pandemic has catalyzed an increasing interest in the best practices for online data collection methods in developmental science. As COVID-19 restrictions persist, remote methods will be paramount to developmental science research. Here, we aimed to contribute to the discussions on online data collection methods. Specifically, we asked whether children who participated in study visits face-to-face and children who participated in study visits online performed differently on both standardized and experimental measures. We examined this question in both typically developing, term-born (TB) children,

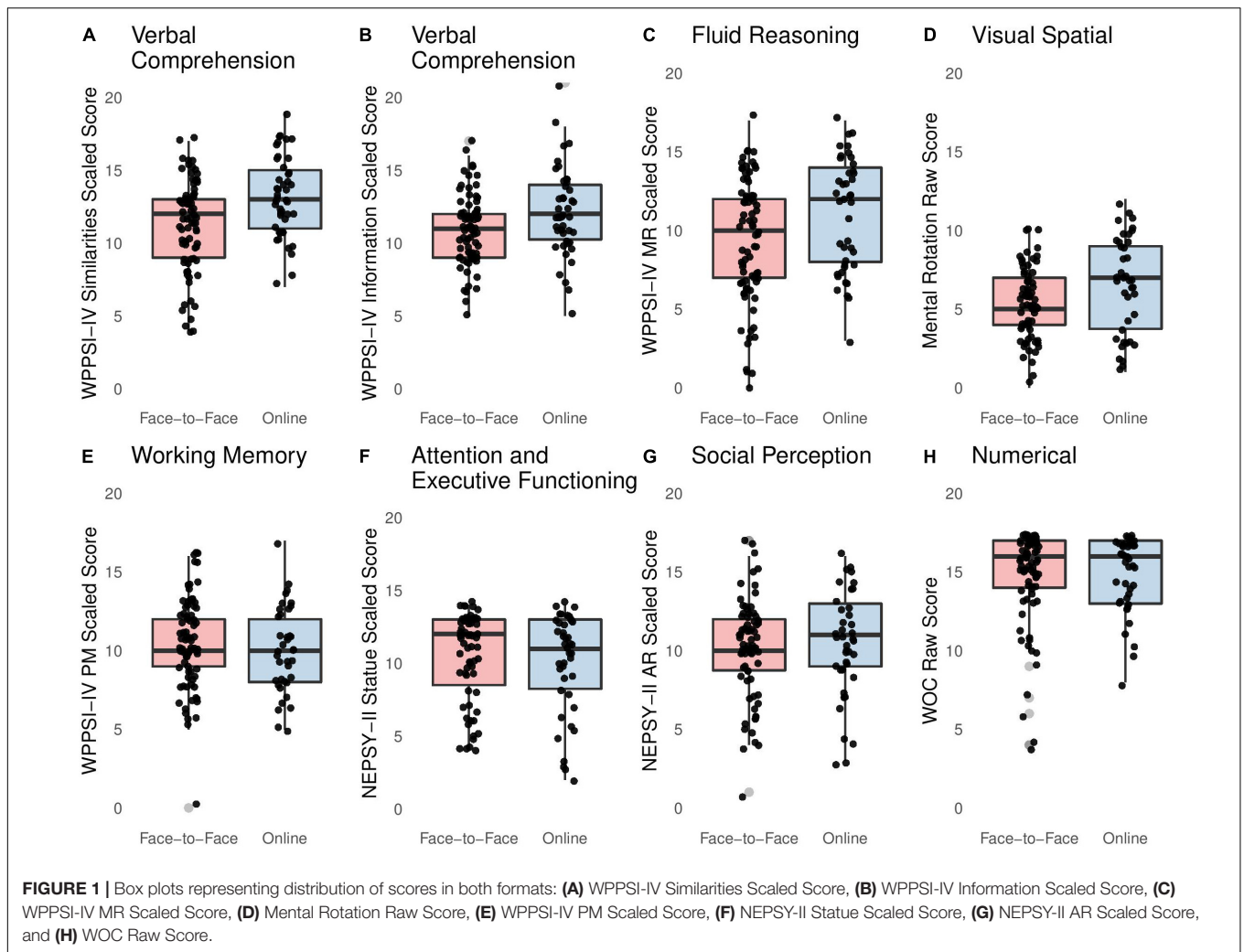


TABLE 3 | Results of mean scores captured in both formats, participants included and excluded, and regression analyses comparing format controlling for parent education and child age at testing.

| | Face-to-face | | | | Online | | | | Regression | | | |
|--|------------------------|----------|------------|--|------------------------|----------|------------|--|-----------------------------|------------------------|----------|--------------------------------|
| | <i>M</i> (<i>SD</i>) | <i>n</i> | <i>Ex.</i> | | <i>M</i> (<i>SD</i>) | <i>n</i> | <i>Ex.</i> | | Beta estimate (<i>SE</i>) | <i>t</i> value | <i>p</i> | Adjusted <i>R</i> ² |
| Verbal comprehension | | | | | | | | | | | | |
| WPPSI-IV similarities | 11.22 (3.15) | 76 | 7 | | 13.18 (2.68) | 45 | 3 | | 1.86 (0.56) | <i>t</i> (117) = 3.33 | <0.01** | 0.10 |
| WPPSI-IV information | 10.92 (2.41) | 77 | 6 | | 12.35 (3.29) | 46 | 2 | | 1.46 (0.50) | <i>t</i> (119) = 2.94 | <0.01** | 0.14 |
| Fluid reasoning | | | | | | | | | | | | |
| WPPSI-IV matrix reasoning | 9.28 (3.93) | 78 | 5 | | 11.02 (3.54) | 42 | 6 | | 1.83 (0.73) | <i>t</i> (116) = 2.50 | 0.01* | 0.04 |
| Visual spatial | | | | | | | | | | | | |
| Mental rotation | 5.37 (2.22) | 70 | 13 | | 6.55 (3.07) | 40 | 8 | | 0.20 (0.46) | <i>t</i> (105) = 0.43 | 0.67 | 0.31 |
| Working memory | | | | | | | | | | | | |
| WPPSI-IV picture memory | 10.48 (2.86) | 79 | 4 | | 10.03 (2.78) | 38 | 10 | | −0.47 (0.56) | <i>t</i> (114) = −0.83 | 0.41 | −0.01 |
| Attention and executive functioning | | | | | | | | | | | | |
| NEPSY-II statue | 10.38 (3.09) | 63 | 20 | | 10.05 (3.41) | 42 | 6 | | −0.36 (0.65) | <i>t</i> (101) = −0.55 | 0.58 | −0.01 |
| Social perception | | | | | | | | | | | | |
| NEPSY-II affect recognition | 10.07 (3.12) | 76 | 7 | | 10.33 (3.35) | 40 | 8 | | 0.30 (0.61) | <i>t</i> (112) = 0.50 | 0.62 | 0.03 |
| Numerical | | | | | | | | | | | | |
| What's on this card | 14.65 (3.03) | 78 | 5 | | 14.93 (2.36) | 41 | 7 | | −0.33 (0.57) | <i>t</i> (114) = −0.58 | 0.56 | 0.06 |

Statistics represent beta estimate, standard error, *T* Value, and *P* value for the Effect of Format. *Ex.*, excluded. **p* < 0.05; ***p* < 0.01.

and in at-risk, preterm-born (PTB) children. We also explored whether parental involvement in the online format related to children's performance.

The finding that children's performance did not vary across the two formats for most of the measures that we administered provides support for the utilization of online data collection methods in developmental science. Here, we provide empirical evidence suggesting that children's performance was not significantly influenced by format on a wide range of cognitive assessments, both standardized and experimental. Our results may help alleviate some of the concerns that researchers have raised about online data collection methods. First, researchers have expressed concern about having less control over the testing environment and having a higher number of distractions, which might lead to a greater portion of data being excluded from online formats. However, we examined this hypothesis, and this was not the case in our sample. It is important to highlight that, in both formats, trained clinical science graduate students audited study visits and excluded data that was thought to be an inaccurate representation of the child's performance. For example, if it was clear that the child was not answering questions due to shyness, that child's data was excluded. Second, there has been concern about differences in sample characteristics between samples that participate face-to-face and samples that participate online, including concerns about differences in demographic characteristics (e.g., parental education) and in health characteristics (e.g., children with special health needs). In our sample, group comparisons did not reveal any significant sociodemographic differences between the two groups. It should be noted that the average income and average education of our sample were generally high, so differences might emerge in a more socioeconomically diverse sample.

Moreover, we were able to recruit PTB children, who are at-risk for academic challenges. Although we were not sufficiently powered to include analyses on interactions between prematurity and format, we conducted exploratory analyses examining such interactions. Reliable bivariate interactions between format and prematurity did not emerge on any of the tests, except for WPPSI-IV MR; PTB children performed lower than TB children in the online format, but not face-to-face format. Thus, overall, format did not appear to greatly undermine preterm children's performance. Future studies that are sufficiently powered should examine interactions between format and risk factors.

While children's performance did not vary by format on most of the assessments, their scores on WPPSI-IV information, similarities, and matrix reasoning did. Average scores on these three subtests were lower for those who participated in the face-to-face format. This finding is inconsistent with findings from a previous study, suggesting equivalence in online and face-to-face scores on information, similarities, and matrix reasoning (Wright, 2020). However, this study included older children and leveraged proctors instead of parents. Thus, it is possible that parents played a role in our findings. However, our analyses showed that children who had more parental involvement and children who had no parental involvement did not differ on information, similarities or matrix reasoning. Another possibility is that the online format lent itself to performing better, due

to children participating in the study in the comfort of their own homes. During face-to-face study visits, children completed testing in an unfamiliar lab while their parents were in another room; during online study visits, children completed testing in their homes while their parents were sitting next to them. While we might expect to see this increased comfort reflected in other tests as well, comfort at the beginning of the study might have differed to a greater extent across the two formats. Information, similarities, and matrix reasoning were administered during the first part of the study visit in both formats. However, in the online format, parents and children played together for 5–10 min prior to the tasks. Although the children were familiarized with the research assistant via play in the lab as well, playing with specifically the parent in the online format could have made children feel more comfortable prior to testing. Finally, it is possible that those who participated online were exposed to factors, whether related to the pandemic or not, that those who participated face-to-face were not. Maybe those who participated online had experiences that benefited their performance on information, similarities, and matrix reasoning. For example, it is possible that those who participated online spent more time interacting with their parents than those who participated face-to-face, due to parents being at home more during the pandemic. Indeed, the verbal skills and acquired knowledge involved in information and similarities, for example, may be sensitive to parental input (Kan et al., 2011; Pace et al., 2016). However, the other tasks included in this study may also benefit from parental input (Clingan-Siverly et al., 2021). Thus, future research should explore whether different types of parental input influenced some cognitive abilities but not others.

Our study has limitations that should be discussed for future studies. First, children may have performed better in the online format due to ordering effects. We did not counterbalance the standardized neurocognitive assessments and experimental tasks for the face-to-face portion of the study. Therefore, we stayed consistent when structuring the online portion of the study and continued with the same order of administration. Future research would benefit from counterbalancing the battery of neurocognitive assessments. These neurocognitive assessments demand a lot of attention and children can easily become fatigued throughout the span of assessments. Second, our combined sample was predominately TB children, compared to PTB children. Future research would benefit from having balanced numbers of TB and PTB children. Third, our combined sample was also predominately of a high socioeconomic status. Future research would benefit from having a more diverse sample, so we could better generalize our findings. Last, due to the small sample size, especially in the online group, we may be underpowered to detect main, and more likely, interaction effects. Thus, we argue that our findings should be replicated in future studies with larger samples.

Taken together, our results suggest that online data collection might be a feasible option for several cognitive measures, for both PTB and TB children. Our results also suggest that, however, online data collection for certain measures, including WPPSI-IV information, similarities, and matrix reasoning should be interpreted with caution. Future research should examine

the mechanisms through which data collection format might influence children's performance. Relevant factors to consider include parental involvement and familiarity with the setting. In addition to our empirical findings, we demonstrated success in recruiting and including several families from lower-resourced rural communities and several families with preterm children to participate online, highlighting the feasibility of including samples with different demographic and health characteristics when using online methods. This not only has implications for research methods but also for providing prevention and intervention services. Our results contribute to the growing body of literature examining differences and similarities across various data collection methods. Online data collection may be a good option for continuing developmental psychology research, for diversifying research samples, and for providing services (e.g., educational services and clinical services) when traditional methods are not available. Our findings can also inform future studies hoping to explore the use of online test administration for educational and clinical purposes, as online methods may be more convenient and accessible for both providers and families.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Iowa Institutional Review

Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

ÖD-L, PN, and FS conceptualized the study and wrote the manuscript. PN, FS, and HL collected the data. ÖD-L and PN analyzed the data. All authors prepared the data for analysis and contributed to the article and approved the submitted version.

FUNDING

This research was supported by NICHD R03HD102449 and the Centers for Disease Control and Prevention of the U.S. Department of Health and Human Services (HHS) as part of a Cooperative Agreement Number (U48 DP006389). FS was supported by the NIHGM grant T32GM108540. The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by CDC/HHS, or the U.S. Government.

ACKNOWLEDGMENTS

The authors would like to thank the families who participated in our study and research assistants who helped with data collection.

REFERENCES

- Allotey, J., Zamora, J., Cheong-See, F., Kalidindi, M., Arroyo-Manzano, D., Asztalos, E., et al. (2018). Cognitive, motor, behavioural and academic performances of children born preterm: a meta analysis and systematic review involving 64 061 children. *BJOG* 125, 16–25. doi: 10.1111/1471-0528.14832
- Bhutia, A. T., Cleves, M. A., Casey, P. H., Cradock, M. M., and Anand, K. J. (2002). Cognitive and behavioral outcomes of school-aged children who were born preterm: a meta-analysis. *JAMA* 288, 728–737. doi: 10.1001/jama.288.6.728
- Brooks, B. L., Sherman, E. M., and Strauss, E. (2009). NEPSY-II: a developmental neuropsychological assessment. *Child Neuropsychol.* 16, 80–101. doi: 10.1080/09297040903146966
- Brydges, C. R., Landes, J. K., Reid, C. L., Campbell, C., French, N., and Anderson, M. (2018). Cognitive outcomes in children and adolescents born very preterm: a meta analysis. *Dev. Med. Child Neurol.* 60, 452–468. doi: 10.1111/dmcn.13685
- Bureau, U. S. C. (2021). Nearly 93% of Households With School-Age Children Report Some Form of Distance Learning During COVID-19. *Census.gov*. Available online at: <https://www.census.gov/library/stories/2020/08/schooling-during-the-covid-19-pandemic.html> (accessed October 15, 2021).
- Centers for Disease Control and Prevention [CDC] (2020). *Preterm Birth*. Available online at: <https://www.cdc.gov/reproductivehealth/maternalinfanthealth/pretermbirth.htm> (accessed April 15, 2021).
- Clingan-Siverly, S., Nelson, P. M., Göksun, T., and Demir-Lira, ÖE. (2021). Spatial thinking in term and preterm-born preschoolers: relations to parent-child speech and gesture. *Front. Psychol.* 12:651678. doi: 10.3389/fpsyg.2021.651678
- EducationData (2021). *Research and Resources to Tackle Critical Issues in Education*. Available online at: <https://educationdata.org/online-education-statistics> (accessed April 15, 2021).
- Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., and Yurovsky, D. (2016). Using tablets to collect data from young children. *J. Cogn. Dev.* 17, 1–17. doi: 10.1080/15248372.2015.1061528
- Garrisi, K., King, C. J., Hillyer, L., and Gaab, N. (2020). *General Recommendations and Guidelines for Remote Assessment of Toddlers and Children*. doi: 10.31219/osf.io/wg4ef
- Kan, K. J., Kievit, R. A., Dolan, C., and van der Maas, H. (2011). On the interpretation of the CHC factor Gc. *Intelligence* 39, 292–302. doi: 10.1016/j.intell.2011.05.003
- Levine, S. C., Huttenlocher, J., Taylor, A., and Langrock, A. (1999). Early sex differences in spatial skill. *Dev. Psychol.* 35, 940–949. doi: 10.1037/0012-1649.35.4.940
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Manning, B. L., Harpole, A., Harriott, E. M., Postolowicz, K., and Norton, E. S. (2020). Taking language samples home: feasibility, reliability, and validity of child language samples conducted remotely with video chat versus in-person. *J. Speech Lang. Hear. Res.* 63, 3982–3990. doi: 10.1044/2020_JSLHR-20-00202
- Morini, G., and Blair, M. (2021). Webcams, songs, and vocabulary learning: a comparison of in-person and remote data collection as a way of moving forward with child-language research. *Front. Psychol.* 12:702819. doi: 10.3389/fpsyg.2021.702819

- Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., and Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra Psychol.* 6:17213. doi: 10.1525/collabra.17213
- Pace, A., Levine, D., Morini, G., Hirsh-Pasek, K., and Golinkoff, R. M. (2016). "The story of language acquisition: from words to world and back again," in *Child Psychology: A Handbook of Contemporary Issues*, 3rd Edn, eds. L. Balter and C. Tamis-LeMonda 43–79.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Snijders, V. E., Bogicevic, L., Verhoeven, M., and van Baar, A. L. (2020). Toddlers' language development: the gradual effect of gestational age, attention capacities, and maternal sensitivity. *Int. J. Environ. Res. Public Health* 17:7926. doi: 10.3390/ijerph17217926
- Su, I. A., and Ceci, S. (2021). "Zoom Developmentalists": home-based videoconferencing developmental research during COVID-19. *PsyArXiv* [Preprint]. doi: 10.31234/osf.io/nvdy6
- Tsega, M., Giantris, K., and Shah, T. (2020). *Essential Social Services are Struggling to Survive the COVID-19 Crisis*. New York, NY: The Commonwealth Fund.
- United Nations Children's Fund (2021). *COVID-19: Schools for more than 168 Million Children Globally have been Completely Closed for Almost a Full Year, says UNICEF*. New York, NY: UNICEF.
- Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence*, 4th Edn. San Antonio, TX: The Psychological Corporation.
- Wolke, D., Johnson, S., and Mendonça, M. (2019). The life course consequences of very preterm birth. *Annu. Rev. Dev. Psychol.* 1, 69–92. doi: 10.1146/annurev-devpsych-121318-084804
- Wright, A. J. (2020). Equivalence of remote, digital administration and traditional, in-person administration of the Wechsler Intelligence Scale for Children, (WISC-V). *Psychol. Assess.* 32, 809–817. doi: 10.1037/pas0000939
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.
- Copyright © 2021 Nelson, Scheiber, Laughlin and Demir-Lira. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Remote Research Methods: Considerations for Work With Children

Michelle M. Shields*, Morgan N. McGinnis and Diana Selmeczy*

Cognitive Development Lab, Department of Psychology, University of Colorado, Colorado Springs, Colorado Springs, CO, United States

OPEN ACCESS

Edited by:

Natasha Kirkham,
Birkbeck, University of London,
United Kingdom

Reviewed by:

Mary M. Flaherty,
University of Illinois
at Urbana-Champaign, United States
Naomi Sweller,
Macquarie University, Australia

*Correspondence:

Michelle M. Shields
mgarci24@uccs.edu
Diana Selmeczy
Diana.Selmeczy@uccs.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 April 2021

Accepted: 27 September 2021

Published: 28 October 2021

Citation:

Shields MM, McGinnis MN and
Selmeczy D (2021) Remote Research
Methods: Considerations for Work
With Children.
Front. Psychol. 12:703706.
doi: 10.3389/fpsyg.2021.703706

The growing shift to online research provides numerous potential opportunities, including greater sample diversity and more efficient data collection. While online methods and recruitment platforms have gained popularity in research with adults, there is relatively little guidance on best practices for how to conduct remote research with children. The current review discusses how to conduct remote behavioral research with children and adolescents using moderated (i.e., real-time interactions between the experimenter and child) and unmoderated (i.e., independent completion of study without experimenter interaction) methods. We examine considerations regarding sample diversity and provide recommendations on implementing remote research with children, including discussions about remote software, study design, and data quality. These recommendations can promote the use of remote research amongst developmental psychologists by contributing to our knowledge of effective online research practices and helping to build standardized guidelines when working with children.

Keywords: remote research, remote research design, remote research software, development, children

INTRODUCTION

Researchers have grown increasingly interested in collecting data using online or remote methodologies. Remote research provides several benefits, such as the potential for quicker data collection and the inclusion of more diverse participant samples (Buhrmester et al., 2011; Dworkin et al., 2016). However, remote methods may also present unique challenges, including difficulties transferring in-person studies to remote formats and the potential for lower quality data due to less direct control over the environmental setting (Bridges et al., 2020; Chmielewski and Kucker, 2020). Previous research using remote methods has mainly been conducted with adults (Paolacci and Chandler, 2014; Lee et al., 2018), and we have a limited understanding of how to best implement remote methodologies with developmental populations (Sheskin et al., 2020). Research conducted with children versus adults can vary substantially, such as differences in instructions and task design (Barker and Weller, 2003). Therefore, it is important to develop appropriate remote research practices that apply to developmental populations. Below we explore remote research methodologies with typically developing child and adolescent populations, focusing on behavioral research in cognitive psychology.

Before assessing the use of remote methods, it is important to note that remote research can be conducted in multiple formats. Unmoderated remote studies utilize online software that allows participants to complete a study individually, at any time, without the presence of a researcher. In contrast, moderated remote studies take place virtually such that researchers interact directly with participants through virtual meeting platforms (e.g., Zoom) and lead participants

through the study procedure in real-time. We will include general considerations regarding both remote unmoderated and moderated methods to help investigators understand the benefits and drawbacks of each format.

BENEFITS OF REMOTE RESEARCH AND COLLECTING REPRESENTATIVE SAMPLES

Diverse samples are critical to developmental research given the large amount of variability that occurs within developmental processes, including cognitive skills (Rowley and Camacho, 2015; Nielsen et al., 2017). Furthermore, developmental processes may be susceptible to environmental effects and vary as a function of ethnicity, socioeconomic status (SES), and geographical location (Bradley and Corwyn, 2002; McCulloch, 2006; Quintana et al., 2006). However, psychological research tends to collect data from homogenous or non-representative samples (Rowley and Camacho, 2015), and this may occur, in part, because most academic research uses in-person or lab-based studies. In-person studies may limit the diversity of research samples due to geographical, temporal, and fiscal restrictions (Rowley and Camacho, 2015; Nielsen et al., 2017; Rhodes et al., 2020). Importantly, remote methods have the potential to overcome some of these limitations by removing the time and costs associated with traveling to a physical research location and allowing individuals to participate at any time (in unmoderated studies). Consistent with this idea, some research shows that both adult and children samples collected through remote research have greater racial and geographical diversity than in-person studies (Birnbaum, 2004; Rhodes et al., 2020).

Despite the potential to increase sample diversity through remote research methods, diversity may still be limited for multiple reasons. For example, internet access, computer access, and technological literacy, which are frequently required to participate in remote studies, are often raised as critical barriers to participation (Kraut et al., 2004; Scott et al., 2017; Grootswagers, 2020). Furthermore, although remote research may decrease the need for travel, having children participate in remote studies may still be time-consuming for parents. For example, parents may need to answer scheduling emails, prescreening forms, or questionnaires, and provide consent or help during the study session. Thus, although remote studies have the potential to increase diversity, there are still limiting factors and future research is needed to determine whether the use of remote research can successfully increase diversity in child samples.

IMPLEMENTING REMOTE RESEARCH STUDIES

Research with children is generally considered more challenging than research with adults because tasks need to be adapted to appropriately match children's language, comprehension, and executive function abilities, and children tend to be more subject to fatigue during participation (Fargas-Malet et al., 2010;

Rollins and Riggins, 2021). Similar to in-person research, including engaging, meaningful, and easy to understand task content (Fargas-Malet et al., 2010; Nussenbaum et al., 2020) is also important when conducting remote research with children. Furthermore, although researchers can remotely collect physiological measures, including eye movements (Scott et al., 2017), most remote work collects behavioral responses. There are some instances when remote research may not be possible, such as when special equipment (e.g., EEG) or highly controlled environmental contexts are required. Below we outline considerations regarding software, experimental design, and data quality for remote behavioral research using typically developing children.

Remote Technology

Remote behavioral research typically requires software that participants can interact with on their own devices (e.g., mobile phones, tablets, laptops, or computers). Several software and online platforms exist to aid researchers in remote data collection (see **Table 1**). A complete summary of available software is beyond the scope of the current review. We suggest researchers examine available software to select the one that best fits their needs. For example, some programs are available through an internet browser (e.g., Qualtrics, Gorilla Experiment Builder), while other programs may require participants to install software on specific operating systems or devices (e.g., Eprime Go). Online software may also vary in its flexibility to implement research designs. For example, Qualtrics is commonly used to collect survey responses but has limited functions for complex coding (e.g., randomization based on multiple variables).

Researchers who work with children and adolescents should also consider participants' development capabilities regarding technology use when designing remote studies. Although more research is needed on children's evolving technological skills, direct observations of children's interaction with technology show that toddlers (Geist, 2012) and infants as young as 15-months-old are able to tap on touch screen devices (Zack et al., 2009; Ziemer et al., 2021). Preschoolers can engage in more complex touch actions such as drag-and-drop (Vatavu et al., 2015). Furthermore, both direct observations and parental reports suggest that 2.5-year-olds begin to use a mouse or keyboard input and 5-year-olds begin to develop basic typing skills with substantial improvements throughout middle childhood (Read et al., 2001; Calvert et al., 2005; Donker and Reitsma, 2007; Kiefer et al., 2015). Therefore, researchers must adopt technological methods that can accommodate the fine-motor skills of their participants, such as using mobile devices or tablets to collect touch input when working with younger children. Researchers should also consider using software that enables video/audio recordings (e.g., Gorilla) or using video conference programs (e.g., Zoom) to collect verbal rather than typed responses for younger participants. Furthermore, children's previous experience with technological devices can also impact research findings (Couse and Chen, 2010; Jin and Lin, 2021), suggesting that researchers should measure children's familiarity with technology as a potential covariate.

TABLE 1 | Comparison of remote software.

| | Gorilla | Inquisit Web | PsyToolkit | EPrime3/Eprime-Go | Qualtrics | Psychstudio | PsychoPy3/PsychoJs |
|------------------------------|--|---|--|---|--|--------------------------|--|
| Remote platform | Web-based | Web-based* -Includes offline feature through application *Requires local download by researcher for customization | Web-based | Web-based* *Requires local download by researcher to create study, may require local download by participant for running study, supported only by Windows OS | Web-based -Includes offline feature through application | Web-based | Web-based* *Requires local download by researcher for creating study and additional software (Pavlovio) to run online |
| Programming language | Typescript (super-set of Javascript) and Handlebars (HTML templating engine) | Similarities to HTML/XML and C-family of languages | Custom scripting language, C- family languages | Custom object- oriented scripting language (E-Basic) | HTML, CSS, and JavaScript | No custom code available | Python |
| Input measures | Mouse Keyboard Audio recording Video recording Mouse- tracking Eye-tracking | Mouse Keyboard Audio recording | Mouse Keyboard | Mouse Keyboard | Mouse Keyboard | Mouse Keyboard | Mouse Keyboard |
| Pricing model | Free to use Pay per participant | License fee | Free | License fee | Free basic version License fee for full version | License Fee | Free |
| Additional supported devices | Mobile Tablet | Mobile Tablet | Mobile Tablet | Tablet | Mobile Tablet | Mobile Tablet | Mobile Tablet |

This table is not an exhaustive list and additional features may be available.

Differences in hardware, software, and response modality may also impact the precision and accuracy of display times, location of stimuli, or response times (Chetverikov and Upravitelev, 2016; Poth et al., 2018). Although remote research software has relatively minimal display and reaction time delays (<100 ms) (Anwyl-Irvine et al., 2020; Bridges et al., 2020), variability exists between browsers, operating systems, and hardware (Garaizar and Reips, 2019; Bridges et al., 2020). Additionally, certain input types (e.g., touch) can result in faster reaction times compared to other input modalities (e.g., mouse) (Woodward et al., 2017; Ross-Sheehy et al., 2021), suggesting it is important to control for input type when measuring reaction times. Critically, general findings may replicate across study methods, with recent research suggesting that response time patterns in children ages 4–12 are similar across remote and in-person studies (Nussenbaum et al., 2020; Ross-Sheehy et al., 2021). Overall, researchers who need highly precise stimuli presentation or response times should instruct participants to use a particular setup during study sessions (e.g., Chrome browser and keyboard), calibrate programs to adjust for the type of operating system and device used, and use within-subjects comparisons or controls (Bridges et al., 2020).

Study Design and Data Quality

Researchers have less control over the experimental environment during remote research, potentially lowering data quality. Remote methods can differ from in-person research in terms of participant engagement (Dandurand et al., 2008), response honesty (Shapiro et al., 2013), and susceptibility to scammers (Dennis et al., 2018). We discuss these factors below and include recommendations on how to overcome some of these challenges.

Task Considerations and Instructions

Remote studies may result in fewer participant–researcher interactions, especially in unmoderated remote research. Although this may be less of a concern in research with adults, the cognitive skills required to independently guide oneself through a task, including self-regulation and language abilities, develop substantially throughout childhood (Montroy et al., 2016; Skibbe et al., 2019). Additionally, infants through preschoolers learn better from in-person interactions than pre-recorded videos (DeLoache et al., 2010; Myers et al., 2017). However, social exchanges that occur virtually in real time (e.g., video chatting) have been shown to be effective even for young children's learning (Strouse and Samson, 2021). Therefore, moderated remote methods where virtual participant–researcher interactions occur may be especially appropriate with younger children. However, unmoderated methods are still possible when additional considerations are used, such as comprehensive instructions, comprehension checks, and parental involvement (Oppenheimer et al., 2009; Kannass et al., 2010; Scott et al., 2017). Furthermore, developmental differences in reading ability can be lessened by using age-appropriate, prerecorded instructions.

Parental involvement may increase during remote relative to in-person studies. For example, parents need to be able to operate and troubleshoot the technological software used for remote research. Because of this, we suggest the use of browser-based

platforms and to limit the use of special software that requires local downloads (see **Table 1**). Furthermore, we suggest that prior to the study session, researchers provide parents with step-by-step instructions on how to use software (see <https://osf.io/wahky/> for guides on using Zoom from our lab) and information on what type of hardware can be used (e.g., mobile phones, tablets, laptops). Critically, due to COVID-19, adults' technological literacy (Sari, 2021) and children's time spent interacting with technology has increased (Drouin et al., 2020; Ribner et al., 2021). These changes have likely made it easier for parents and children to implement basic functions in video conferencing platforms (e.g., video/audio communication and screen sharing) and other software. However, we recommend that researchers add approximately 10 min of additional time during study sessions to troubleshoot any technological issues and prepare to reschedule sessions if needed.

Researchers may also want to intentionally direct parental involvement during data collection. Parental support and scaffolding can be helpful, especially when working with younger children. Recent research shows that during remote sessions having parents input responses for children ages 4–10 results in similar findings as in-person studies (Ross-Sheehy et al., 2021), providing some evidence that parental involvement can be used successfully during remote research. However, researchers may often want to prevent unwanted parental involvement (e.g., additional unmonitored explanations, biasing of responses) or require children to input their own responses, especially if accurate response times are needed. As children learn to communicate independently, they may be less likely to need parental intervention, with research suggesting children as young as 4 years of age can independently input their responses during remote research (Vales et al., 2021). To limit parental involvement during data collection, researchers can read instructions to children or use pre-recorded audio or videos (Rhodes et al., 2020). During moderated sessions, researchers could also share their screen and input children's responses or have children share their screen and monitor children's behaviors while children input their own responses. We also recommend that researchers communicate to parents the importance of children's independent responses. Additionally, we suggest researchers collect feedback from both children and parents on any issues that may have come up during the study, such as cheating or asking for parental help.

Increasing Attention and Motivation

Lack of participant attention during remote research, including increased distractions and decreased motivation, can lower data quality (Zwarun and Hall, 2014; Finley and Penningroth, 2015). Participants are also more likely to experience distractions in natural settings outside of a research laboratory, and these distractions can lead to different findings than those observed during lab-based studies (Kane et al., 2007; Varao-Sousa et al., 2018). Furthermore, children and adolescents have greater difficulty ignoring irrelevant information (Davidson et al., 2006; Garon et al., 2008), and therefore environmental distractions may be more likely to impact remote research with developmental populations. In addition to distractions, it is possible that

participants may be less motivated during remote studies and rapidly complete tasks or provided unvaried answers (Litman et al., 2015; Ahler et al., 2020).

Several methods have been found to reduce participant inattention during remote research with adults. Attention checks, including trap questions (e.g., regardless of your true preference select “Movies”) can be used to flag inattentive participants (Liu et al., 2009; Hunt and Scheetz, 2018). Comprehension checks (e.g., what are your instructions for this task?) can also be used to help researchers ensure that participants understand and are attentive to the task. Researchers can then use predetermined criteria for removing participants based on responses to these questions to improve overall data quality (Dworkin et al., 2016; Jensen-Doss et al., 2021). When working with children, trap questions (e.g., answer this question by pressing the blue button) and comprehension checks (e.g., select the option that shows what you will be doing in the study) that require specific age-appropriate responses can also be included to assess and remove inattentive participants. Moderated studies with children may be inherently more engaging and therefore less susceptible to low levels of attention and motivation (Dandurand et al., 2008), but researchers can still continue to directly monitor, address, and note participant attention. Additionally, shorter, engaging tasks may improve attention during remote research, including the use of animations, child-friendly stimuli, and frequent breaks (Barker and Weller, 2003; Rhodes et al., 2020).

Limiting Cheating

Another concern that can affect data quality is the honesty of participants’ responses. Participants may be more likely to answer dishonestly on tasks completed in the absence of researcher supervision (Lilleholt et al., 2020). The percentage of adults that cheat during online studies can vary (e.g., ranging from 24 to 41% according to Clifford and Jerit, 2014), but research suggests that most adult participants answer honestly when encouraged to do so (Corrigan-Gibbs et al., 2015). However, cheating behaviors may differ when working with developmental populations, with research suggesting younger children ages 8–10 cheat more frequently than older children ages 11–16 during in-person studies (Evans and Lee, 2011; Ding et al., 2014).

Several methods have been shown to decrease cheating behaviors. Simple interventions such as honesty reminders (e.g., “please answer honestly”) and honesty checks were found to decrease cheating behaviors in adults (Clifford and Jerit, 2014; Corrigan-Gibbs et al., 2015) and children (Heyman et al., 2015), and these types of interventions can easily be included in either moderated or unmoderated remote research. During unmoderated sessions, researchers could also minimize cheating by recording participants or taking periodic video captures of participants. During moderated sessions, researchers can monitor participants via video and screen-sharing, and verbally intervene if cheating behaviors are observed. In our own remote research, we have found that nearly all families consent to video recording (>99%) during moderated sessions, suggesting video monitoring is a potentially feasible solution to help mitigate cheating (see <https://osf.io/hrp4y/> for our consent documents). Finally, task designs may need to be altered to mitigate cheating, particularly during memory tasks during which cheating can

easily occur (e.g., writing down to-be-learned material). To minimize cheating, memory researchers can avoid stimuli that can easily be labeled and instead use abstract, similar, or difficult to label stimuli (e.g., scenes, fractals), limit encoding time and require participants to complete an additional task during encoding (e.g., mouse-click on the presented stimuli), or use incidental encoding designs where participants are unaware that an upcoming memory test will occur.

Avoiding Bots and Scammers

Remote studies with minimal researcher interaction may be at risk for compromised data quality due to information-security threats (Teitcher et al., 2015). Previous research with adults has highlighted information-security issues and offers potential solutions (Ahler et al., 2020; Chmielewski and Kucker, 2020). For example, automated computer program responses (i.e., bots) tend to differ from human responses and consist of atypical text formats, grammatically incorrect text responses, or responses that directly copy prompt text (Chmielewski and Kucker, 2020). Therefore, bots are relatively easy to flag and remove. Implementing bot checks (e.g., captchas) and collecting participant screening questions can also decrease bots (Jones et al., 2015; Kennedy et al., 2020). However, scammers may be particularly problematic for remote developmental research, especially during unmoderated designs. Scammers often fabricate responses to receive compensation (Chandler and Paolacci, 2017), including falsely claiming to be of a key demographic (e.g., an adult claiming to be a child). To alleviate some of these issues, researchers can utilize prescreening questions and check for consistencies in responses, such as asking about a child’s age repeatedly and in multiple formats (e.g., DOB, numeric age) or requesting specific information relevant to identifying your targeted population (e.g., asking a parent to describe a recent moment they were proud of their child) (Jones et al., 2015). Email requests to participate in a research study can also be monitored for potential scammers. Instances of strange email addresses, rapidly incoming email inquiries, and inquiries consisting of unusual responses (grammatical errors, copied text, etc.), may further indicate potential scammers. Ultimately, moderated studies may be the most effective at reducing scammers as direct participant-researcher interactions can easily ensure human participants are completing the study.

CONCLUSION

As remote research becomes more common, understanding its benefits and limitations is increasingly important. Above, we outlined several considerations for implementing remote research with children and adolescents, including information about participant samples, remote technologies, study design, and data quality. Although future work is needed to better understand how remote research differs between children and adults, and which methods are most effective for children, the provided recommendations contribute to building a guideline for effective and reliable remote research with developmental populations.

AUTHOR CONTRIBUTIONS

MS contributed to the development and writing of the manuscript. MM contributed to the literature review

process and editing of the manuscript. DS contributed to the development and revision of the manuscript. All authors read and approved the submitted version of the manuscript.

REFERENCES

- Ahler, D. J., Roush, C. E., and Sood, G. (2020). The Micro-Task Market for Lemons: Data Quality on Amazon's Mechanical Turk. Meeting of the Midwest Political Science Association. Available online at: <http://www.gsood.com/research/papers/turk.pdf>
- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., and Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behav. Res. Methods* 53, 1407–1425. doi: 10.3758/s13428-020-01501-5
- Barker, J., and Weller, S. (2003). "Is it fun?" developing children centered research methods. *Int. J. Sociol. Soc. Pol.* 23, 33–58. doi: 10.1108/01443330310790435
- Birnbaum, M. H. (2004). Human research and data collection via the internet. *Annu. Rev. Psychol.* 55, 803–832. doi: 10.1146/annurev.psych.55.090902.141601
- Bradley, R. H., and Corwyn, R. F. (2002). Socioeconomic status and child development. *Annu. Rev. Psychol.* 53, 371–399. doi: 10.1146/annurev.psych.53.100901.135233
- Bridges, D., Pitiot, A., MacAskill, M. R., and Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *PeerJ* 8, 1–29. doi: 10.7717/peerj.9414
- Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980
- Calvert, S. L., Rideout, V. J., Woolard, J. L., Barr, R. F., and Strouse, G. A. (2005). Age, ethnicity, and socioeconomic patterns in early computer use: a national survey. *Am. Behav. Sci.* 48, 590–607. doi: 10.1177/0002764204271508
- Chandler, J. J., and Paolacci, G. (2017). Lie for a dime: when most prescreening responses are honest but most study participants are impostors. *Soc. Psychol. Personal. Sci.* 8, 500–508. doi: 10.1177/1948550617698203
- Chetverikov, A., and Upravitelev, P. (2016). Online versus offline: the Web as a medium for response time data collection. *Behav. Res. Methods* 48, 1086–1099. doi: 10.3758/s13428-015-0632-x
- Chmielewski, M., and Kucker, S. C. (2020). An MTurk Crisis? Shifts in data quality and the impact on study results. *Soc. Psychol. Personal. Sci.* 11, 464–473. doi: 10.1177/1948550619875149
- Clifford, S., and Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *J. Exp. Polit. Sci.* 1, 120–131. doi: 10.1017/xps.2014.5
- Corrigan-Gibbs, H., Gupta, N., Northcutt, C., Cutrell, E., and Thies, W. (2015). Detering cheating in online environments. *ACM Trans. Comput.-Hum. Interact.* 22, 1–23. doi: 10.1145/2810239
- Couse, L. J., and Chen, D. W. (2010). A tablet computer for young children? Exploring its viability for early childhood education. *J. Res. Technol. Educ.* 43, 75–98. doi: 10.1080/15391523.2010.10782562
- Dandurand, F., Shultz, T. R., and Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behav. Res. Methods* 40, 428–434. doi: 10.3758/BRM.40.2.428
- Davidson, M. C., Amso, D., Anderson, L. C., and Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia* 44, 2037–2078. doi: 10.1016/j.neuropsychologia.2006.02.006
- DeLoache, J. S., Chiong, C., Sherman, K., Islam, N., Vanderborght, M., Troseth, G. L., et al. (2010). Do babies learn from baby media? *Psychol. Sci.* 21, 1570–1574. doi: 10.1177/0956797610384145
- Dennis, S., Goodson, B. M., and Pearson, C. (2018). Online worker fraud and evolving threats to the integrity of MTurk data: a discussion of virtual private servers and the limitations of IP-Based screening procedures. *Behav. Res. Account.* 2019, 1–55. doi: 10.2139/ssrn.3233954
- Ding, X. P., Omrin, D. S., Evans, A. D., Fu, G., Chen, G., and Lee, K. (2014). Elementary school children's cheating behavior and its cognitive correlates. *J. Exp. Child Psychol.* 121, 85–95. doi: 10.1016/j.jecp.2013.12.005
- Donker, A., and Reitsma, P. (2007). Young children's ability to use a computer mouse. *Comput. Educ.* 48, 602–617. doi: 10.1016/j.compedu.2005.05.001
- Drouin, M., McDaniel, B. T., Pater, J., and Toscos, T. (2020). How parents and their children used social media and technology at the beginning of the COVID-19 pandemic and associations with anxiety. *Cyberpsychol. Behav. Soc. Network.* 23, 727–736. doi: 10.1089/cyber.2020.0284
- Dworkin, J., Hessel, H., Gliske, K., and Rudi, J. H. (2016). A comparison of three online recruitment strategies for engaging parents. *Fam. Relat.* 65, 550–561. doi: 10.1111/fare.12206
- Evans, A. D., and Lee, K. (2011). Verbal deception from late childhood to middle adolescence and its relation to executive functioning skills. *Dev. Psychol.* 47, 1108–1116. doi: 10.1037/a0023425
- Fargas-Malet, M., McSherry, D., Larkin, E., and Robinson, C. (2010). Research with children: methodological issues and innovative techniques. *J. Early Childhood Res.* 8, 175–192. doi: 10.1177/1476718x09345412
- Finley, A. J., and Penningroth, S. L. (2015). "Online versus in-lab: pros and cons of an online prospective memory experiment," in *Advances in Psychology Research*, Vol. 113, eds A. M. Columbus (Hauppauge, NY: Nova Science Publishers, Inc.), 135–162.
- Garaizar, P., and Reips, U. D. (2019). Best practices: two Web-browser-based methods for stimulus presentation in behavioral experiments with high-resolution timing requirements. *Behav. Res. Methods* 51, 1441–1453. doi: 10.3758/s13428-018-1126-4
- Garon, N., Bryson, S. E., and Smith, I. M. (2008). Executive function in preschoolers: a review using an integrative framework. *Psychol. Bull.* 134, 31–60. doi: 10.1037/0033-2909.134.1.31
- Geist, E. A. (2012). A qualitative examination of two year-olds interaction with tablet based interactive technology. *J. Instruct. Psychol.* 39, 26–35. <https://link.gale.com/apps/doc/A303641377/HRCA?u=colosprings&sid=summon&xid=ae8c6f6e>.
- Grootswagers, T. (2020). A primer on running human behavioural experiments online. *Behav. Res. Methods* 52, 2283–2286. doi: 10.3758/s13428-020-01395-3
- Heyman, G. D., Fu, G., Lin, J., Qian, M. K., and Lee, K. (2015). Eliciting promises from children reduces cheating. *J. Exp. Child Psychol.* 139, 242–248. doi: 10.1016/j.jecp.2015.04.013
- Hunt, N. C., and Scheetz, A. M. (2018). Using MTurk to distribute a survey or experiment: methodological considerations. *J. Inform. Syst.* 33, 43–65. doi: 10.2308/isys-52021
- Jensen-Doss, A., Patel, Z. S., Casline, E., Mora Ringle, V. A., and Timpano, K. R. (2021). Using mechanical turk to study parents and children: an examination of data quality and representativeness. *J. Clin. Child Adolescent Psychol. [Online ahead of print]* 1–15. doi: 10.1080/15374416.2020.1815205
- Jin, Y. R., and Lin, L. Y. (2021). Relationship between touchscreen tablet usage time and attention performance in young children. *J. Res. Technol. Educ.* 1–10. doi: 10.1080/15391523.2021.1891995
- Jones, M. S., House, L. A., and Gao, Z. (2015). Respondent screening and revealed preference axioms: testing quarantining methods for enhanced data quality in web panel surveys. *Public Opin. Q.* 79, 687–709. doi: 10.1093/poq/nfv015
- Kane, M. J., Brown, L. H., McVay, J. C., Silvia, P. J., Myin-Germeyns, I., and Kwapil, T. R. (2007). For whom the mind wanders, and when: an experience-sampling study of working memory and executive control in daily life. *Psychol. Sci.* 18, 614–621. doi: 10.1111/j.1467-9280.2007.01948.x
- Kannass, K. N., Colombo, J., and Wyss, N. (2010). Now, pay attention! the effects of instruction on children's attention. *J. Cogn. Dev.* 11, 509–532. doi: 10.1080/15248372.2010.516418
- Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., and Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Sci. Res. Methods* 8, 614–629. doi: 10.1017/psrm.2020.6
- Kiefer, M., Schuler, S., Mayer, C., Trumpp, N. M., Hille, K., and Sachse, S. (2015). Handwriting or Typewriting? The influence of pen- or keyboard-based writing

- training on reading and writing performance in preschool children. *Adv. Cogn. Psychol.* 11, 136–146. doi: 10.5709/acp-0178-7
- Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., and Couper, M. (2004). Psychological research online: report of board of scientific affairs' advisory group on the conduct of research on the internet. *Am. Psychol.* 59, 105–117. doi: 10.1037/0003-066X.59.2.105
- Lee, Y. S., Seo, Y. W., and Siemsen, E. (2018). Running behavioral operations experiments using Amazon's mechanical turk. *Product. Operat. Manag.* 27, 973–989. doi: 10.1111/poms.12841
- Lilleholt, L., Schild, C., and Zettler, I. (2020). Not all computerized cheating tasks are equal: a comparison of computerized and non-computerized versions of a cheating task. *J. Econ. Psychol.* 78:102270. doi: 10.1016/j.joep.2020.102270
- Litman, L., Robinson, J., and Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behav. Res. Methods* 47, 519–528. doi: 10.3758/s13428-014-0483-x
- Liu, D., Sabbagh, M. A., Gehring, W. J., and Wellman, H. M. (2009). Neural correlates of children's theory of mind development. *Child Dev.* 80, 318–326. doi: 10.1111/j.1467-8624.2009.01262.x
- McCulloch, A. (2006). Variation in children's cognitive and behavioural adjustment between different types of place in the British National Child Development Study. *Soc. Sci. Med.* 62, 1865–1879. doi: 10.1016/j.socscimed.2005.08.048
- Montroy, J. J., Bowles, R. P., Skibbe, L. E., McClelland, M. M., and Morrison, F. J. (2016). The development of self-regulation across early childhood. *Dev. Psychol.* 52, 1744–1762. doi: 10.1037/dev0000159
- Myers, L. J., LeWitt, R. B., Gallo, R. E., and Maselli, N. M. (2017). Baby FaceTime: can toddlers learn from online video chat? *Dev. Sci.* 20:e12430. doi: 10.1111/desc.12430
- Nielsen, M., Haun, D., Kärtner, J., and Legare, C. H. (2017). The persistent sampling bias in developmental psychology: a call to action. *J. Exp. Child Psychol.* 162, 31–38. doi: 10.1016/j.jecp.2017.04.017
- Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., and Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychol.* 6, 17213. doi: 10.1525/collabra.17213
- Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* 45, 867–872. doi: 10.1016/j.jesp.2009.03.009
- Paolacci, G., and Chandler, J. (2014). Inside the turk: understanding mechanical turk as a participant pool. *Curr. Direct. Psychol. Sci.* 23, 184–188. doi: 10.1177/0963721414531598
- Poth, C. H., Foerster, R. M., Behler, C., Schwanecke, U., Schneider, W. X., and Botsch, M. (2018). Ultrahigh temporal resolution of visual presentation using gaming monitors and G-Sync. *Behav. Res. Methods* 50, 26–38. doi: 10.3758/s13428-017-1003-6
- Quintana, S. M., Chao, R. K., Cross, W. E., Hughes, D., Gall, S. N., Aboud, F. E., et al. (2006). Race, ethnicity, and culture in child development: contemporary research and future directions. *Child Dev.* 77, 1129–1141. doi: 10.1111/j.1467-8624.2006.00951
- Read, J., MacFarlane, S., and Casey, C. (2001). "Measuring the usability of text input methods for children," in *People and Computers XV—Interaction Without Frontiers*, eds A. Blandford, J. Vanderdonck, and P. Gray (London: Springer). doi: 10.1007/978-1-4471-0353-0_35
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493.
- Ribner, A. D., Coulanges, L., Friedman, S., and Libertus, M. E. (2021). Screen time in the COVID Era: international trends of increasing use among 3- to 7-Year-Old Children. *J. Pediatr.* doi: 10.1016/j.jpeds.2021.08.06
- Rollins, L., and Riggins, T. (2021). Adapting event-related potential research paradigms for children: considerations from research on the development of recognition memory. *Dev. Psychobiol.* 63:e22159. doi: 10.1002/dev.22159
- Ross-Sheehy, S., Reynolds, E., and Eschman, B. (2021). Unsupervised online assessment of visual working memory in 4- to 10-Year-Old Children: array size influences capacity estimates and task performance. *Front. Psychol.* 12:692228. doi: 10.3389/fpsyg.2021.692228
- Rowley, S. J., and Camacho, T. C. (2015). Increasing diversity in cognitive developmental research: issues and solutions. *J. Cogn. Dev.* 16, 683–692. doi: 10.1080/15248372.2014.976224
- Sari, M. K. (2021). The impacts of Covid-19 pandemic in term of technology literacy usage on students learning experience. *J. Sos. Humaniora JSH* 43–51. doi: 10.12962/j24433527.v0i0.8348
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (Part 2): assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/opmi_a_00001
- Shapiro, D. N., Chandler, J., and Mueller, P. A. (2013). Using mechanical turk to study clinical populations. *Clin. Psychol. Sci.* 1, 213–220. doi: 10.1177/2167702612469015
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Skibbe, L. E., Montroy, J. J., Bowles, R. P., and Morrison, F. J. (2019). Self-regulation and the development of literacy and language achievement from preschool through second grade. *Early Childhood Res. Q.* 46, 240–251. doi: 10.1016/j.ecresq.2018.02.005
- Strouse, G. A., and Samson, J. E. (2021). Learning from video: a meta-analysis of the video deficit in children ages 0 to 6 years. *Child Dev.* 92, e20–e38. doi: 10.1111/cdev.13429
- Teitcher, J. E. F., Bockting, W. O., Bauermeister, J. A., Hoefer, C. J., Miner, M. H., and Klitzman, R. L. (2015). Detecting, preventing, and responding to "Fraudsters" in internet research: ethics and tradeoffs. *J. Law Med. Ethics* 43, 116–133. doi: 10.1111/jlme.12200
- Vales, C., Wu, C., Torrance, J., Shannon, H., States, S. L., and Fisher, A. V. (2021). Research at a distance: replicating semantic differentiation effects using remote data collection with children participants. *Front. Psychol.* 12:697550. doi: 10.3389/fpsyg.2021.697550
- Varao-Sousa, T. L., Smilek, D., and Kingstone, A. (2018). In the lab and in the wild: how distraction and mind wandering affect attention and memory. *Cogn. Res.: Principles Implications* 3:42. doi: 10.1186/s41235-018-0137-0
- Vatavu, R. D., Cramariuc, G., and Schipor, D. M. (2015). Touch interaction for children aged 3 to 6 years: experimental findings and relationship to motor skills. *Int. J. Hum. Comput. Stud.* 74, 54–76. doi: 10.1016/j.ijhcs.2014.10.007
- Woodward, J., Shaw, A., Aloba, A., Jain, A., Ruiz, J., and Anthony, L. (2017). "Tablets, tabletops, and smartphones: cross-platform comparisons of children's touchscreen interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, (New York, NY: ACM), 5–14.
- Zack, E., Barr, R., Gerhardstein, P., Dickerson, K., and Meltzoff, A. N. (2009). Infant imitation from television using novel touch screen technology. *Br. J. Dev. Psychol.* 27, 13–26. doi: 10.1348/026151008x334700
- Ziemer, C. J., Wyss, S., and Rhinehart, K. (2021). The origins of touchscreen competence: examining infants' exploration of touchscreens. *Infant Behav. Dev.* 64:101609.
- Zwarun, L., and Hall, A. (2014). What's going on? Age, distraction, and multitasking during online survey taking. *Comput. Hum. Behav.* 41, 236–244. doi: 10.1016/j.chb.2014.09.041

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Shields, McGinnis and Selmecky. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



OPEN ACCESS

Edited by:

Sho Tsuji,
The University of Tokyo, Japan

Reviewed by:

Nayeli Gonzalez-Gomez,
Oxford Brookes University,
United Kingdom
Martin Zettersten,
Princeton University, United States
Toben H. Mintz,
University of Southern California,
United States

*Correspondence:

Mireia Marimon
marimon@uni-potsdam.de

†ORCID:

Mireia Marimon
orcid.org/0000-0001-7401-7532
Andrea Hofmann
orcid.org/0000-0002-2639-5499
João Veríssimo
orcid.org/0000-0002-1264-3017
Claudia Männel
orcid.org/0000-0003-0678-4697
Angela D. Friederici
orcid.org/0000-0002-6328-865X
Barbara Höhle
orcid.org/0000-0002-9240-6117
Isabell Wartenburger
orcid.org/0000-0001-5116-4441

†These authors share first authorship

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 01 July 2021

Accepted: 13 October 2021

Published: 03 November 2021

Citation:

Marimon M, Hofmann A,
Veríssimo J, Männel C, Friederici AD,
Höhle B and Wartenburger I (2021)
Children's Learning of Non-adjacent
Dependencies Using a Web-Based
Computer Game Setting.
Front. Psychol. 12:734877.
doi: 10.3389/fpsyg.2021.734877

Children's Learning of Non-adjacent Dependencies Using a Web-Based Computer Game Setting

Mireia Marimon^{1*†}, Andrea Hofmann^{1,2†}, João Veríssimo^{1,3†}, Claudia Männel^{4,5†},
Angela D. Friederici^{4†}, Barbara Höhle^{1†} and Isabell Wartenburger^{1†}

¹ Cognitive Sciences, Department of Linguistics, University of Potsdam, Potsdam, Germany, ² Early Childhood Education Research, University of Applied Sciences, Potsdam, Germany, ³ School of Arts and Humanities, University of Lisbon, Lisbon, Portugal, ⁴ Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany, ⁵ Department of Audiology and Phoniatrics, Charité – Universitätsmedizin Berlin, Berlin, Germany

Infants show impressive speech decoding abilities and detect acoustic regularities that highlight the syntactic relations of a language, often coded *via* non-adjacent dependencies (NADs, e.g., *is singing*). It has been claimed that infants learn NADs implicitly and associatively through passive listening and that there is a shift from effortless associative learning to a more controlled learning of NADs after the age of 2 years, potentially driven by the maturation of the prefrontal cortex. To investigate if older children are able to learn NADs, Lammertink et al. (2019) recently developed a word-monitoring serial reaction time (SRT) task and could show that 6–11-year-old children learned the NADs, as their reaction times (RTs) increased then they were presented with violated NADs. In the current study we adapted their experimental paradigm and tested NAD learning in a younger group of 52 children between the age of 4–8 years in a remote, web-based, game-like setting (*whack-a-mole*). Children were exposed to Italian phrases containing NADs and had to monitor the occurrence of a target syllable, which was the second element of the NAD. After exposure, children did a “Stem Completion” task in which they were presented with the first element of the NAD and had to choose the second element of the NAD to complete the stimuli. Our findings show that, despite large variability in the data, children aged 4–8 years are sensitive to NADs; they show the expected differences in RTs in the SRT task and could transfer the NAD-rule in the Stem Completion task. We discuss these results with respect to the development of NAD dependency learning in childhood and the practical impact and limitations of collecting these data in a web-based setting.

Keywords: non-adjacent dependencies, rule learning, web-based, implicit learning, serial reaction time (SRT) task, SRT

INTRODUCTION

To acquire their native language, infants not only have to learn the words but also the rule-based relations between the individual words, which make up the syntax of that language. Some of these grammatical rules, known as non-adjacent dependencies (NADs), consist of statistically reliable relationships between two speech elements separated by intervening elements. An example from

English is the morphological relation between an auxiliary *is* and a verb suffix *-ing* in *My brother is dancing*. The ability to extract and track NADs from speech is crucial for language acquisition (Kidd and Arciuli, 2016; Lany and Shoaib, 2020). Infants have been shown to be able to learn NADs from passive listening (e.g., Gómez, 2002; Gómez and Maye, 2005; Friederici et al., 2011; Marchetto and Bonatti, 2013; for a review see Wilson et al., 2018). However, this learning seems to be hindered under certain conditions, for example, if the variability of the intervening element is low (Gómez, 2002), if the NADs are embedded in complex passages (Santelmann and Jusczyk, 1998), or if the stream does not contain any mark for segmentation (Marchetto and Bonatti, 2013). Although NAD learning might continue to be difficult for children in their second year of life it becomes more and more sophisticated over development (e.g., across phonological word boundaries, generalization to new contexts, more complex patterns, etc.). For instance, infants from 17 months of age show NAD learning even when the discrimination required was extremely subtle (e.g., *pel kicey rud* vs. *pel kicey jic*) (Gómez and Maye, 2005, for English-learning infants) and even when the auxiliary and verb suffix crossed a phonological phrase boundary (van Heugten and Shi, 2010, for French-learning infants). Later, at 19 months of age, infants can recognize NADs over two intervening syllables (Höhle et al., 2006, for German-learning infants). Adults have been shown to be successful NAD learners when tested under passive listening conditions with behavioral methods (Uddén et al., 2012; Frost and Monaghan, 2016; Wang et al., 2019). However, evidence from electrophysiological and neuroimaging studies using identical materials and task settings have shown differences between adults' and infants' NAD learning from passive listening. These studies are outlined in the next paragraph.

Friederici et al. (2011) showed that already 4-month-old German-learning infants can track NADs in an unfamiliar natural language, namely Italian. The authors measured event-related potentials (ERPs) while infants heard Italian sentences consisting of a noun phrase followed by an NAD (e.g., *Il fratello sta cantando*, the brother *is* singing). The stimuli alternated familiarization (learning) and subsequent test phases which contained the familiarized NADs and violations of the NADs (e.g., **Il fratello sta cantare*, **the brother is sing*; **means agrammatical*). Results showed a broad positive-going ERP component in response to the NAD violations indicating that infants could discriminate between familiarized and violated NADs after a short period of passive listening, and thus were sensitive to NADs. This P600-like positivity was similar to the response of adult native Italian speakers, but differed from German speaking adults, who showed an N400 effect that was taken as evidence for lexical, rather than syntactic processing (Mueller et al., 2009). Only after a prolonged time of exposure (Citron et al., 2011) or explicit instructions (Mueller et al., 2012) German speakers' brain responses showed a similar pattern as the ones of the native speakers.

For NAD learning in early childhood, Mueller et al. (2019) reported that 2-year-olds, but not 4-year-olds showed ERP markers of rule learning from passive listening. van der Kant et al. (2020) further narrowed down the period of this developmental

change and showed that NAD learning undergoes a qualitative change between 2 and 3 years of age. Their results indicated learning of NADs *via* passive listening for children at the age of 2 years, but not at the age of 3 years. In line with these findings, it has been proposed that the ability to learn implicitly (i.e., without instruction and/or feedback) from passive listening declines from early infancy to later childhood (Skeide and Friederici, 2016). The question arises as to whether the capacity for a more associative bottom-up learning from passive listening ends abruptly around the age of 3 years or whether it might gradually be replaced by a more top-down, controlled learning mechanism. Paul et al. (2020) investigated this transition of NAD learning and collected ERP data from children between 1 and 3 years of age. Using the same experimental paradigm as in the above cited studies (Mueller et al., 2009; Citron et al., 2011; Friederici et al., 2011) they observed that the amplitude of the ERP effect of NAD learning decreased linearly with age suggesting a gradual decrease of NAD learning from passive listening. Importantly, Paul et al. (2020) argued on the one hand for a developmental shift, presumably influenced by maturation of the prefrontal cortex (PFC) and other neuronal circuits (Skeide and Friederici, 2016), but on the other hand also proposed that children's knowledge and entrenchment of their native language has an influence on the changes in their learning outcomes. According to Skeide and Friederici (2016), when maturation has reached a certain degree, top-down control increasingly takes effect, which in turn inhibits associative bottom-up learning mechanisms, also limiting the ability to learn NADs under passive listening. In line with this idea, Friederici et al. (2013) demonstrated that in a passive listening experiment, in which adults' left prefrontal region was downregulated with a cathodal transcranial direct current stimulation, they showed a late positivity for violated NADs similar to infants, indicating associative learning. In the control sham-condition, adults showed the lexical N400-like component as in Mueller et al. (2009). The developmental shift from more associative to more controlled learning mechanisms thus seems to be related to the development of the PFC functions.

So far, there has been little evidence for this developmental decline in NAD learning from behavioral paradigms. One of the possible hurdles is that behavioral data collection in children is limited: grammaticality judgments (e.g., two-alternative-forced-choice task, 2AFC), reaction time (RT) and reflection-based measures (Isbilen et al., 2017), typically used with adult participants, are challenging for children (Lammertink et al., 2019). Lammertink et al. (2019) developed a promising methodological setting to examine children's NAD learning behavior by adapting a serial reaction time (SRT) task combined with a word-monitoring task for children aged 5;9 to 8;6 years (see also López-Barroso et al., 2016). In this children-friendly game setting (in the lab), participants were introduced to two little monkeys on a computer screen and were asked to help the monkeys gather bananas, while they were exposed to an artificial language string containing items with and without NADs. Children were then asked to press a button as fast as possible if they heard a specific target syllable (Version 1: target "lut" in "tep X lut;" Version 2: target "mip" in "sot X mip;" X stands for 72 variable elements between the NADs) and another non-target

button if the target syllable was not presented. During the initial “learning blocks,” children reacted faster over time in response to the target- and non-target syllables. But crucially, they slowed down in the so-called “reversal block” (in the present study we name it “disruption block”), in which the second element of the NAD (“lut” in Version 1 and “mip” in Version 2) was not preceded by the first element of the learned NAD (“tep” or “sot,” respectively), but by a novel syllable. After that, children were presented with the correct NADs (“recovery block”) and again, were faster in their responses. These results showed that children were sensitive to the NADs, because the first element predicted the second element, resulting in this specific RT pattern. In addition, children completed a grammaticality judgment task (2AFC), a more explicit task that tests for abstraction and transfer of the NAD to a new setting. However, in this task children performed only at chance level. The authors argued that, although 2AFC measures are widely used to test NAD learning, the required degree of metalinguistic or explicit knowledge may have influenced the judgments, possibly invalidating their use with children. This is corroborated by Bialystok (1986), who indicates that metalinguistic skills are acquired and mastered not until the age of 7 years.

In the present study, we aim to replicate and extend Lammertink et al.'s (2019) study by examining the ability to learn NADs in younger children (4–8-year-olds) in a web-based active SRT learning task with natural language stimuli (adapted from Friederici et al., 2011). As in Lammertink et al. (2019), we employed an active task, as we asked participants to interact and actively press buttons in response to the presented stimuli. In addition, we adapted the 2AFC task of Lammertink et al. (2019) to test the abstraction and transfer of the internal rule to a new setting. Thus, our study aims to address (1) whether implicit learning of NADs in 4–8-year-old children can be captured by means of RTs in an active web-based task, (2) whether NAD learning can additionally be measured by means of a Stem Completion (SC) task, and (3) whether children's NAD learning (in either task) is modulated by age. Importantly, our work differs from Lammertink et al. (2019) in the following aspects:

Age Range

Existing literature on NAD learning has either mainly explored early infancy up to the age of 4 years (e.g., Gómez, 2002; Mueller et al., 2019; Paul et al., 2020; van der Kant et al., 2020), older children (e.g., 6–8-year-olds; Lammertink et al., 2019), or adults (e.g., Frost and Monaghan, 2016; Arnon, 2019) with data on children between 4 and 6 years still missing. However, there is evidence that a developmental shift happens gradually between 2 and 4 years of age (Mueller et al., 2018; Paul et al., 2020; van der Kant et al., 2020). To systematically investigate the question of how NAD learning trajectories unfold and what influences the magnitude of NAD learning beyond the age of 3 years, we collected behavioral data in children from 4 to 8 years of age. So far, to the best of our knowledge, no empirical data of SRT learning measures for NAD learning or the combination of RTs and response accuracy measures exists for children across the whole the age range of 4–8 years (besides the partial overlap with Lammertink et al., 2019).

Stem Completion Task

Lammertink et al. (2019) included a 2AFC task in which children heard pairs of utterances and had to decide which of the two utterances was most familiar to the artificial language heard in the previous word-monitoring task (e.g., “tep X lut” or “tep X mip”). However, children did not exceed chance level in this task. In our study we included an SC task instead. Children heard only the first part of the NAD without the final element (e.g., “sta cant-”) and were asked to decide which ending would fit best (i.e., target syllable, “ando” or non-target syllable “are”) by clicking on the respective button on the keyboard. Thus, our task is still a decision task, but it includes two main deviations: firstly, children are not presented with the alternative options and then forced to choose between these two, but rather must decide on the best possible completion of the stimulus from two possible “hidden” options (“are” or “ando”) without hearing the “complete” stimulus. Secondly, with this approach we did not have to create a new task environment with completely new instructions, but the children had to continue behaving in a similar manner as in the SRT task, that is, monitoring the target syllable. However, we consider this task more explicit than the SRT task, because participants need to access the previously learned underlying NAD rule.

Natural Stimuli

We used natural language stimuli (adapted from Friederici et al., 2011) instead of artificial syllable strings or phrases (e.g., “tep X lut,”; Lammertink et al., 2019). Experiments using artificial languages have received criticism in recent years in terms of ecological validity (Yang, 2004; for a review see Erickson and Thiessen, 2015). Compared to natural language, artificial languages are relatively simple in their acoustic properties and contain less variability that defines rhythm and stress characteristics.

Web-Based Study

Finally, due to the worldwide pandemic situation (COVID-19), our study was fully run at home on an Internet browser instead of in the laboratory. Whereas web-based data collection in adults is extensively and successfully used and several well-established RT effects have been replicated in web-based research (Crump et al., 2013; Simcox and Fiez, 2014), there are only a few recent web-based studies with children and infants (Scott et al., 2017; Nussenbaum et al., 2020; Rhodes et al., 2020; Bambha and Casasola, 2021; Vales et al., 2021). Recent studies collecting RTs with adults and children have shown little to no difference between laboratory-based and web-based samples (de Leeuw and Motz, 2016; Hilbig, 2016; Bridges et al., 2020; Nussenbaum et al., 2020; Morini and Blair, 2021; Silver et al., 2021; Vales et al., 2021) as well as no big differences between browsers (e.g., Chrome and Internet Explorer) or experiment builders (e.g., Pavlovia and Gorilla) (Kochari, 2019; Anwyl-Irvine et al., 2020; Sauter et al., 2020). In addition, it is difficult to clearly state whether potential differences may be any greater than the difference between two laboratory-based collected samples (Nussenbaum et al., 2020).

Furthermore, our study was unmoderated, which means that there was no interaction with the experimenter and children needed minimal assistance from parents. The feasibility of collecting web-based data with children and young infants in unmoderated studies has recently been demonstrated (Scott et al., 2017; for a discussion of the advantages and challenges, see Rhodes et al., 2020; Bambha and Casasola, 2021). Web-based data collection allows to collect larger sample sizes which leads to increased statistical power (Brand and Bradley, 2012; Nussenbaum et al., 2020).

Based on the previous literature, our hypotheses are that all children should be able to learn NADs within the word-monitoring SRT task (learning from passive listening). Specifically, we expected a training effect, a disruption effect and a recovery effect. The **training effect** would be confirmed if children's RTs decreased through the first exposure learning blocks, in which only correct (i.e., to-be-learned) NADs are presented. A **disruption effect** would be confirmed if, as in Lammertink et al. (2019), RTs increased during the disruption block in which children are presented with violated NADs (according to what they have learned in the previous learning blocks). Finally, the expected **recovery effect** would be confirmed if RTs decreased again after shifting back to the correct, to-be-learned NADs (recovery block; same stimuli as in the initial learning blocks). In the subsequent SC task, children had to apply the NAD rule learned during the SRT task. Thus, in this task we test whether the implicitly learned rule of the SRT task can be transferred by the participants to a more explicit knowledge. In addition, a positive correlation between both tasks would indicate that better NAD-learners can better extract and transfer the underlying NAD rule to a different task demand. Finally, we expected that age would modulate the three effects in the SRT task and the accuracy in the SC task. Since the ability to learn from passive listening might decline due to the increased top-down control of the more mature PFC (Friederici et al., 2013; Skeide and Friederici, 2016), older children may show less sensitivity to the NADs in our stimuli than younger children.

MATERIALS AND METHODS

Participants

Overall, 91 monolingual German-speaking children fully completed the first part of the experiment (SRT task). Twenty-three additional children started the study and quit before the first task was finished and therefore their data were not included in the analysis. From the sample of 91 children, 37 children responded randomly on the two buttons [i.e., showed at-chance performance in the word-monitoring task of the SRT part (see section "Data Preprocessing")]. In line with the exclusion criteria of Lammertink et al. (2019), data of these children were excluded from the analysis. Two further participants were excluded because they did the task twice. Two additional datasets were excluded only from the SC task analysis due to incomplete data. In the final sample, a total of 52 children were included (27 girls and 25 boys; age range: 3;7–8;04 years; mean = 6.21 years,

SD = 1.12 years)¹ (see **Table 1**). Participants were recruited through the BabyLab database and the internet portal *Kinder Schaffen Wissen*² and received a compensation of five Euros if requested. Before starting the study, parents reported no speech or hearing disorders for their children and no daily exposure to a Romance language (French, Italian, Spanish, Romanian, or Portuguese) and gave their consent. Ethical approval was obtained from the Ethics Committee of the University of Potsdam (EA 43/2018).

Stimuli

All stimuli (adapted from Friederici et al., 2011) were recorded by a native Italian speaker and consisted of short grammatical and ungrammatical Italian utterances of the form AXB. Here, the A-element represents the first element of the NAD and refers to two different auxiliaries depending on the experimental version: an Italian verb auxiliary (*sta*) or an Italian modal verb (*può*). The middle element X was a variable Italian verb stem. Each verb stem was morphologically marked and contained one of two Italian suffixes (B-element *-are* or *-ando*, depending on the version). Grammatical and ungrammatical NAD stimuli were generated by combining each auxiliary with each suffix and cross-splicing (see Friederici et al., 2011, for a detailed description; see stimuli used here in **Table 2**). Thus, the NAD stimuli contained a monosyllabic A-element (*sta* or *può*) followed by one monosyllabic verb stem (one out of 32 different X-elements), followed by a bisyllabic B-element (*are* or *ando*) (e.g., *sta cantando* and *può cantare*).

The B-element (*are* or *ando*) was the target syllable for the word-monitoring task. The target syllables and the corresponding NADs were counterbalanced across participants. A total of 4 experimental versions were created and counterbalanced across participants to control for any intrinsic biases and saliency toward the native Italian grammatical dependencies (see **Table 3** for all combinations). Thus, half of the participants learned the NADs *sta-X-ando* and *può-X-are*; and for half of them *ando* was the target syllable, for the other half *are* was the target syllable. The other half of the participants learned the NADs *può-X-ando* and *sta-X-are*; and for half of them *ando* was the target, for the other half *are* was the target. For the sake of simplicity, we explain the tasks for only one of these four different lists below (Version 3, see **Figure 1**).

Out of the 32 different Italian verb stems (X-elements), 24 were randomly selected and present in both NADs (e.g., *sta-X-ando* and *può-X-are*) and appeared in a different order in each block. Therefore, in the SRT task, children were presented with these verb stems twice in the learning blocks and the recovery block and once in the disruption block (hence called familiar verb stems). The SC task contained eight trials with familiar verb stems and eight trials with the

¹For reasons of privacy data protection, parents were asked to provide the year and month of their child's birth before starting the experiment. In the analysis a date specification was compiled by setting the numeric of the birthday date to the 15th of each month. Thus, it is possible that children's age deviates from the actual age by a maximum of 15 days. In addition, parents were asked to create a code that would serve as the ID for the data analysis. This ID was only linked to the raw data and could not be traced back to the actual participant.

²www.kinderschaffenwissen.de

TABLE 1 | Summary of participants' characteristics that were included in the data analysis.

| Total | Age in years ^a | Handedness (total) | Other languages (total) | Time to finish study in minutes ^a |
|-------|-----------------------------------|----------------------------------|--|--|
| 27 | 6.05 (1.03, range = 4.01–7.37) | Right (23) Left (2) NA (2) | Female English (2) Finnish (1) Polish (1) | 19.94 (6.88, range = 15.83–30.94) |
| 25 | 6.34 (1.19, range = 3.70–8.04) | Right (23) Both (2) | Male Russian (2) | 23.10 (7.40, range = 16.83–46.99) |

^aMean (SD, range).**TABLE 2** | Stimuli and stimulus specifications.

| NAD-stimuli | | | | |
|--------------------------|-----------------|---------------------------|---------------|---|
| Regular, grammatical | | sta X _i _ando | | A ₁ X _i B ₂ |
| Italian version | | ↑ NAD ↑ | | ↑ NAD ↑ |
| | | può X _i _are | | A ₂ X _i B ₁ |
| | | ↑ NAD ↑ | | ↑ NAD ↑ |
| Irregular, ungrammatical | | *sta X _i _are | | *A ₁ X _i B ₁ |
| Italian version | | ↑ NAD ↑ | | ↑ NAD ↑ |
| | | *può X _i _ando | | *A ₂ X _i B ₂ |
| | | ↑ NAD ↑ | | ↑ NAD ↑ |
| Specifications | | | | |
| Auxiliaries | A ₁ | sta | to be | gerund, first person singular |
| | A ₂ | può | to be able to | modal, first person singular |
| Suffixes | B ₁ | are | is X_ing | infinitive form |
| | B ₂ | ando | can X_Ø | progressive form |
| Example verb stem | XB ₁ | cantare | to sing | infinitive |
| | XB ₂ | cantando | singing | progressive |

*means agrammatical.

remaining novel verb stems that participants did not hear during the SRT task.

Each stimulus contained a silent pause of 20 ms at the beginning and a pause between auxiliary and verb stem (pause in *può* stimuli: mean = 259 ms, min = 148 ms, max = 350 ms; pause in *sta* stimuli: mean = 261 ms, min = 142 ms, max = 361 ms), but no pause between verb stem and suffix. **Table 4** contains all verbs with the respective suffix used in the study. The average trial length for *sta X_ando* trials was 1550 ms (SD = 66 ms), for *sta X_are* trials it was 1450 ms (SD = 70 ms), for *può X_ando* trials it was 1410 ms (SD = 61 ms), and for *può X_are* trials it was 1320 ms (SD = 55 ms). There was no significant difference in the B-element onset (*ando* and *are*) across the stimuli ($p = 0.90$). All auditory instructions for the game were recorded by a female native speaker of German.

Serial Reaction Time Task

The word-monitoring SRT task included a practice block, an exposure phase consisting of two learning blocks followed by a disruption block and a recovery block (**Figure 1**). The practice block consisted of six trials and was fully repeated if children responded incorrectly in more than two trials. Each learning

block consisted of 48 NAD trials (24 target and 24 non-target trials). The disruption block after the learning blocks consisted of 24 trials (12 violated target and 12 violated non-target trials) and was followed by a recovery block of 48 trials (24 target and 24 non-target trials). Every 24 trials children received feedback on the number of stars (correct responses) collected so far.

The stimuli presented in the different blocks were counterbalanced across four different trial types: *target trial*, *non-target trial*, *violated target trial*, and *violated non-target trial*. A target trial contained the target syllable that children were asked to monitor during the experiment by pressing the button when hearing it (e.g., *are*). The non-target trial was therefore determined by the absence of the target syllable: if the target trial contained the target syllable *are*, the non-target trial contained the non-target syllable *ando*. For example, the target syllable in Version 3 was *are*. Hence, a target trial in Version 3 was *sta-X-are* and a non-target trial was *può-X-ando*. These trials were presented during both the learning blocks and the recovery block. In the disruption block, the trials presented contained only “disrupted” NADs (violated target trials and violated non-target trials). In these types of trials, the dependency between the first and the

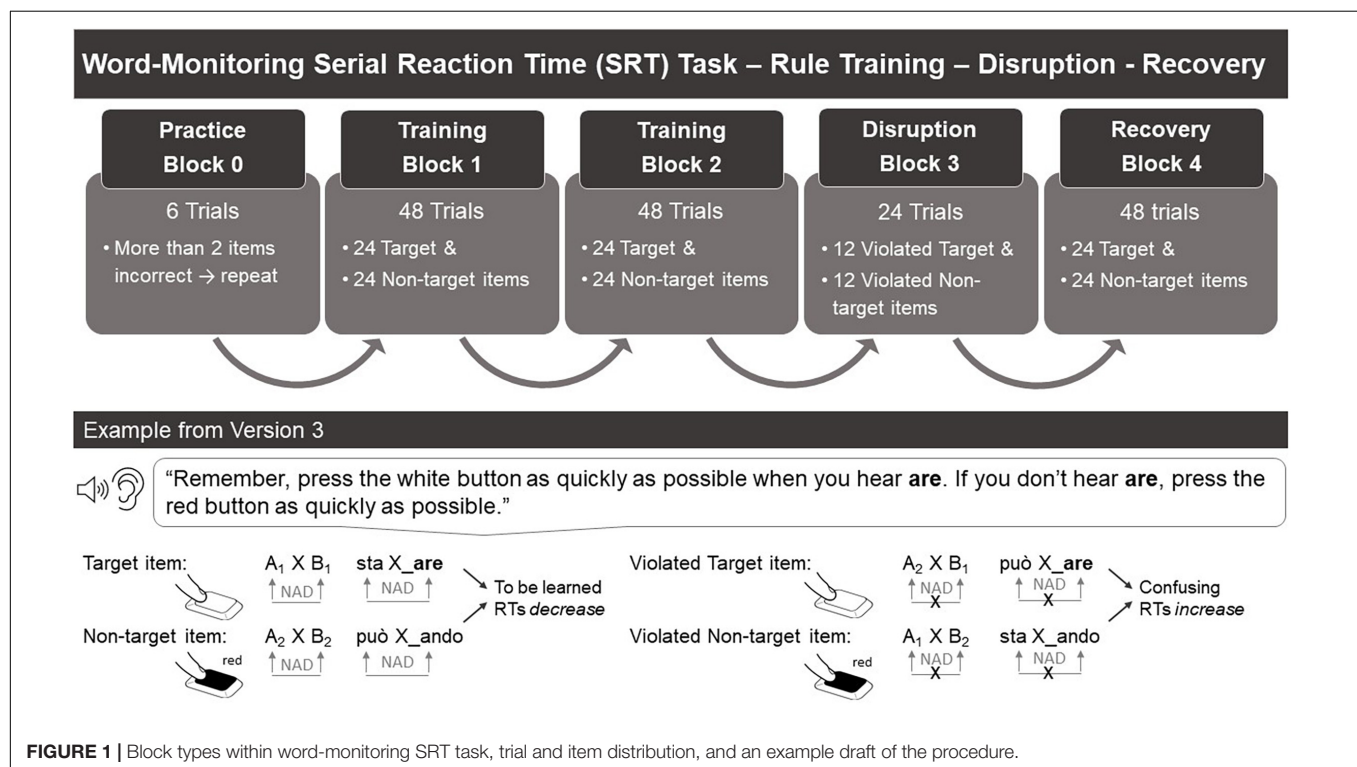
TABLE 3 | Trial types and their use within blocks and each version of the word-monitoring SRT task.

| Word-monitoring SRT task | | | | | |
|--------------------------|---------------------|--|--|--|--|
| Block | Trial type | Version 1 | Version 4 | Version 2 | Version 3 |
| | | Target word <i>ando</i> | | Target word <i>are</i> | |
| Learning and recovery | Target | può X _i <i>_ando</i> A ₂ X _i B ₂ △ | sta X _i <i>_ando</i> A ₁ X _i B ₂ ○ | può X _i <i>_are</i> A ₂ X _i B ₁ □ | sta X _i <i>_are</i> A ₁ X _i B ₁ ★ |
| | Non-target | sta X _i <i>_are</i> A ₁ X _i B ₁ ★ | può X _i <i>_are</i> A ₂ X _i B ₁ □ | sta X _i <i>_ando</i> A ₁ X _i B ₂ ○ | può X _i <i>_ando</i> A ₂ X _i B ₂ △ |
| Disruption | Target violated | sta X _i <i>_ando</i> A ₁ X _i B ₂ ○ | può X _i <i>_ando</i> A ₂ X _i B ₂ △ | sta X _i <i>_are</i> A ₁ X _i B ₁ ★ | può X _i <i>_are</i> A ₂ X _i B ₁ □ |
| | Non-target violated | può X _i <i>_are</i> A ₂ X _i B ₁ □ | sta X _i <i>_are</i> A ₁ X _i B ₁ ★ | può X _i <i>_ando</i> A ₂ X _i B ₂ △ | sta X _i <i>_ando</i> A ₁ X _i B ₂ ○ |

Instruction

“Press the white button as quickly as possible when you hear *ando*. If you do not hear *ando*, press the red button as quickly as possible.”

“Press the white button as quickly as possible when you hear *are*. If you do not hear *are*, press the red button as quickly as possible.”

**FIGURE 1 |** Block types within word-monitoring SRT task, trial and item distribution, and an example draft of the procedure.

second NAD element was violated. For example, a violated target trial in Version 3 was *può-X-are* and a violated non-target trial was *sta-X-ando*. Thus, the violated target trials still contained the target syllable that participants had to monitor (i.e., *are* in our example from Version 3). Therefore, during the SRT task, the children had always to monitor the

same target syllable, which was assigned to the same button throughout the experiment.

Stem Completion Task

Stimuli used in this task consisted only of the first element of the NAD and the verb stem (AX-element, e.g., *può cant-*) and did

TABLE 4 | Overview of the 32 verb stems with the respective suffix combinations, used within stimuli.

| Infinitive | Gerund |
|--|--|
| amare, andare | amando, andando |
| bagnare, ballare, bussare | bagnando, ballando, bussando |
| cantare, cercare, chiamare, cullare | cantando, cercando, chiamando, cullando |
| danzare | danzando |
| entrare | entrando |
| filmare, fischiare | filmando, fischiando |
| gelare, gettare, giocare, girare, graffiare, gridare | gelando, gettando, giocando, girando, graffiando, gridando |
| lodare | lodando |
| mangiare, mostrare | mangiando, mostrando |
| ornare | ornando |
| pagare, pappare, passare, pensare, picchiare | pagando, pappando, passando, pensando, picchiando |
| stirare, suonare | stirando, suonando |
| tirare | tirando |
| volare | volando |

Each verb stem was combined once with *sta* and once with *può*, depending on the experimental version.

not contain the second element of the NAD (B-element, *ando* or *are*), which was cut at zero-crossing from each utterance using the software *Praat* (Boersma and Weenink, 2018). All edited auditory stimuli for the SC task were checked by 17 different adult raters unaware of the experimental procedure to evaluate whether the missing suffix could somehow be derived from the stimuli (e.g., because of subtle coarticulation differences). The raters listened to each stimulus individually and selected through a questionnaire whether *-are* or *-ando* was a better fit at the end. All rater assignments were at chance level and response biases due to inherent stimulus characteristics could therefore be ruled out ($N = 32$ possible correct assignments, $p = 0.5$ probability of success, range of chance level between 12 and 20 correct assignments, actual ratings: between 14 and 19 correct assignments). In the SC task, the participants had to select the correct completion of the NAD by pressing on the target button or non-target button, respectively. For example, in Version 3, in which children were asked to monitor the target syllable *are*, a target trial consisted of *sta cant-*. The children would need to press the target button, as *are* would be the correct answer according to the learned NAD. A non-target trial consisted of *può cant-* and the children would need to press the non-target button, as *ando* would be the correct answer according to the previously learned NAD. The SC task consisted of 16 trials (8 target and 8 non-target) and included feedback (stars) only at the end of the task.

Procedure

The experiment was accessed *via* a link on the BabyLab page of the University of Potsdam and was programmed and deployed on the web-based *LabVanced* software³ (Finger et al., 2017). Before

³The *LabVanced*'s server is located in Germany, does not store any personal data, and follows the EU data privacy regulations (GDPR). More information can be found at the website <https://www.labvanced.com/docs/geninfo/security/>.

starting the experiment, parents were asked to prepare a white and a red button by placing a white sticker on the P key and a red one on the Q key (under the assumption that parents with children at home have paper, colored pencils, and sticky tape at hand). At the beginning of the experiment, parents were asked to fill in a short questionnaire about their child (month and year of birth, sex, handedness, input in other languages, and speech and hearing impairment). After that, parents were requested to test and adjust the volume on the speakers or headphones. Parents were asked not to get involved with the experiment, but to stay close by. After the parental questionnaire, the experiment started and children were introduced to two different cartoon moles, Mali and his brother Max, who invited the child to play a catching game with them (*whack-a-mole*; Nissen and Bullemer, 1987; Qian et al., 2016). For this, children were instructed to listen very carefully because they would listen to phrases from a secret language (the NADs) which sometimes included a specific target syllable and sometimes not. Children were instructed to press the white target button (right hand) as soon as they heard the specific target syllable and the red non-target button (left hand) otherwise. Children were also told that they had to answer questions about the secret language at the end. The instruction was then briefly repeated so that the target syllable (*ando* or *are*) and the corresponding required responses were remembered before a practice block of the SRT task started. RTs were recorded as the dependent variable. At the beginning of each trial, the two moles appeared next to each other and the audio containing a single NAD phrase (e.g., *sta cantando*) was directly played. The child could press the button anytime from stimulus onset. As soon as the child pressed one of the two buttons, the feedback appeared, which depended on the accuracy of the child's response. If the child responded correctly (i.e., pressed the target button if the target syllable was present or pressed the non-target button if the target syllable was not present), the mole on the respective side was caught with a net. As an additional reward, a star appeared and a sound was played (positive feedback). If the child responded incorrectly (i.e., pressed the target button if the target syllable was not present or pressed the non-target button if the target syllable was present), only an empty net appeared in the middle between the two moles and no sound was played. After the child responded, the next trial started with the two moles and a new trial containing an NAD phrase. Pressing the button was self-paced, there was no trial timeout and therefore no null responses were recorded. After the SRT task ended, the instructions for the SC task were explained to the participants. Children were told that they would hear phrases from the secret language again but without an ending. They were asked to choose the best ending (either *are* or *ando*) with the same button press procedure as before and were told that they would receive a star reward at the very end. They were asked to guess if they were unsure. Response accuracy was collected as the dependent variable. A trial in the SC task started the same way as in the SRT task. However, the feedback differed: a red or a white circle appeared on the mole according to the participants' response and independently of the accuracy of the answer.

All children had to perform both tasks immediately after each other, preferably without a pause. At the very end,

parents were asked to indicate whether the children used headphones during the whole experiment. On average, children took 21.6 min to complete the experiment ($SD = 7.32$, range = 15.83–46.99 min). Each participant used their own laptop or desktop⁴ and participants were asked to wear headphones if available. The caregiver of 12 children reported that their child wore headphones throughout the entire experiment and three did so only for the SRT task; the others reported that they used loudspeakers. All collected metrics were provided by *LabVanced* as a downloadable csv file. The experiment, in digital JSON format, along with all scripts used, can be found at our Open Science Framework project page.

DATA ANALYSIS

Data Preprocessing

Following Lammertink et al. (2019), we included only data of those participants in the analysis who were able to follow the instruction of the word-monitoring task and did not respond randomly using the two buttons in the two learning blocks and the recovery block in the SRT task. Hence, an above-chance performance in the word-monitoring task was considered an indication of adequate task compliance. Monitoring the target syllable (i.e., whether the syllable appeared in a sentence) was therefore considered a “secondary” task, which was a relatively easy and cognitively low demanding compared to the main task that consisted in implicitly learning the internal structure of the NADs (see section “Participants”). In the final sample only correct target word monitoring responses were analyzed (78.29% of total number of trials). In addition, three criteria were applied to determine outliers and exclude individual RT data points. First, RTs lower or equal to 200 ms were removed (1.9% of total number of trials), because RTs up to 200 ms from stimulus onset may be too fast to reflect the processes of interest, as they correspond to the approximate duration to plan and execute an adequate motoric response (Dahan et al., 2019). Secondly, since there was no timeout for trials, RTs above 7000 ms were considered long RT outliers based on visual inspection of the data (e.g., Ratcliff, 1993; Baayen and Milin, 2010), as they are more likely not revealing any information about the underlying linguistic processing, and therefore they were removed (2.3% of total number of measures). This specific cut-off was chosen *post hoc* after observing the large variability of the data. Finally, RTs that were 2.5 SD above or below the mean RT for the corresponding target type (target, non-target, violated target, and violated non-target) of the same participant in the same block were removed as well (2.3% of total number of trials). A total percentage of 5.2% of individual RTs were excluded from further analysis based on the described criteria. The final dataset contained 6335 observations, all for correct responses only, distributed over four blocks (first learning block: 1676 out of 1840 observations; second learning block: 1826 out of 1953 observations; disruption block: 917 out of 981 observations; and recovery block: 1916 out of 2065 observations).

⁴All operating systems were allowed. All Internet browsers were allowed, except Safari.

In the SC task, responses were coded as correct or incorrect (accuracy). A correct response (coded as 1) was possible in two ways: (a) if children pressed the target button in a target trial, deciding that the target syllable would be the best fit for completion of the NAD or (b) if children pressed the non-target button in a non-target trial, deciding that the non-target syllable would be the best fit for completion of the NAD. In all other cases the response was incorrect (coded as 0).

Statistical Analysis

We based our analyses on Lammertink et al. (2019), who kindly provided their scripts on OSF.⁵ Both RTs and accuracy data were analyzed in R (R Core Team, 2017) using (generalized) linear mixed-effects models (*lme4* package; Bates et al., 2015). Confidence intervals were calculated using the profile method (provided within *lme4* package), odds ratios and probabilities were calculated following the script of Lammertink et al. (2019) and *p*-values were obtained by loading the *lmerTest* package in R before fitting the model (Kuznetsova et al., 2017). All corresponding figures were generated using the *ggplot2* package (Wickham, 2009). The raw RTs of the final dataset were log-transformed. This was determined by an assessment of the best Box–Cox power transform, a procedure that allows selecting the appropriate data transformation that normalizes the residuals of the statistical models (Box and Cox, 1964; Kliegl et al., 2009; Kowarik, 2019).

Serial Reaction Time Task

We employed a linear mixed-effects model in which log RT was the dependent outcome variable in the model. In the model, *Block* was entered as a fixed effect with four levels: first learning block, second learning block, disruption block, and recovery block. To obtain information about the three effects of interest we used successive difference contrasts that allowed us to directly test the difference between condition means of neighboring block levels. With the generated contrasts, the difference between two successive blocks is tested while condition means for the other block levels is ignored (Schad et al., 2020). The corresponding comparison between mean (log-scaled) RTs made it possible to confirm the presence of the following effects: first, a **training effect**, which contained the difference between the first learning block and second learning block (coded as first learning block = -0.75 , second learning block = $+0.25$, and remaining *Block* levels = $+0.25$). Second, a **disruption effect**, which contained the difference between the second learning block and the disruption block (coded as second learning block = -0.5 , disruption block = $+0.5$, first learning block = -0.5 , and recovery block = $+0.5$). Finally, a **recovery effect** was determined through the difference between the disruption block and the recovery block (coded as disruption block = -0.25 , recovery block = $+0.75$, and remaining blocks = -0.25). *Targetness* (target and non-target) was entered as a fixed effect and coded as a sum-to-zero contrast (difference of the level means between target, coded $+0.5$, and non-target, coded -0.5 , see Schad et al., 2020). *Age* (in years and months) was centered and included as

⁵<https://osf.io/bt8ug/>

a continuous variable. The model contained random intercepts by-subject (*Subject*) and by-item (*Item*) and random slopes by-subject for the main effects of *Targetness* and *Block*, as well as for the interaction between *Targetness* and *Block*, and a random slope by-item for *Age*. Since we modeled fixed effects for all predictors with sum-to-zero contrasts, it allowed us to estimate the respective coefficients as overall effects across the levels of all other predictors, defining the intercept term as the grand mean across all predictor levels. The model structure was selected prior to data collection and was based on Lammertink et al.'s (2019) approach with deviations in the contrast coding and without *Version* as a predictor. To determine a sensible random-effects structure we used a backward-selection heuristic (Matuschek et al., 2017) based on the AIC criterion (Akaike, 1998) to arrive at a model that included the random effects' structure justified by the data without losing goodness-of fit and without losing power to detect fixed effects or substantially increase Type I error rates. Note that the resulting model is highly complex and may therefore lack power to detect any of the interaction effects, especially three-way interactions.

Stem Completion Task

We employed a generalized linear mixed-effects model (mixed-effects logistic regression model) in which accuracy was the dependent outcome variable. We fitted a model to estimate whether children scored better in items with a familiar verb stem compared to items with a novel verb stem (*Familiarity*), and whether children scored better on target trials compared to non-target trials (*Targetness*). The model included two binary predictors, each interacting with *Age* as a covariate. *Familiarity* and *Targetness* were included as fixed factors and both were coded with sum-to-zero contrasts (verb stem: familiar +0.5, novel -0.5; item: target +0.5, non-target -0.5). *Age* (in years and months) was centered and included as a continuous variable. The model contained only by-subject (*Subject*) random intercepts and slopes for the main effect of *Targetness*. The parsimonious random effects' structure was derived by means of a backward-selection heuristic (Matuschek et al., 2017), as in the SRT task model. Finally, we computed a Pearson's correlation coefficient to determine the relationship between the SRT and the SC task. The outcome of the SRT task was calculated by subtracting a child's average log-transformed RT in the disruption block from his/her log-transformed RT average in the second training block and recovery block (i.e., disruption peak). The SC score was the number of correct responses from each child.

RESULTS

Serial Reaction Time Task

We tested whether children's RTs showed the three effects of interest (training effect, disruption effect, and recovery effect). **Figure 2** shows the mean RTs across the blocks according to *Targetness* (target and non-target). The complete output of the model is provided in **Table 5**. The results indicated a statistically significant effect for responses collapsed across both target types for two effects of interest: the training effect and the

disruption effect. The **training effect** in the model output shows that children were 156.22 ms (back-transformed from model estimates) faster in the second learning block compared to the first one [$t = -2.74$; $p = 0.006$; 95% CI (-0.13, -0.02)]. The **disruption effect** was indicated by a 103.56 ms increase in RTs in the disruption block compared to the second learning block [$t = +2.04$; $p = 0.04$; 95% CI (0.00, 0.09)]. The **recovery effect** was not statistically significant: the mean RTs on log scale in the recovery block were 83.81 ms shorter than in the disruption block [$t = -1.75$; $p = 0.08$; 95% CI (-0.09, 0.00)]. In addition, the output of the model showed a main effect of *Targetness*: children were faster responding to target items than to non-target items [-44.37 ms, $t = -2.31$; $p = 0.02$; 95% CI (-0.04, 0.00)]. Finally, we found a main effect of *Age* [$t = -2.66$; $p = 0.01$; 95% CI (-0.16, -0.02)], suggesting that overall older children responded faster than younger children (-186.91 ms on average), and an interaction between *Age* and *Targetness* [$t = -2.02$; $p = 0.04$; 95% CI (-0.03, 0.00)]. Finally, we found a main effect of *Age* [$t = -2.66$; $p = 0.01$; 95% CI (-0.44, -0.07)], suggesting that overall older children responded faster than younger children (-186.91 ms on average), and an interaction between *Age* and *Targetness* [$t = -2.02$; $p = 0.04$; 95% CI (-0.09, -0.00)]. We divided the group into three subgroups according to age and we observed three different types of behavior: younger children seemed to be slower for target items compared to non-target items, the middle age subgroup seemed to show no difference in their responses when presented with either target type, and older children seemed to be faster for targets compared to non-target items. However, *Age* did not significantly modulate the effects of interest (i.e., training effect, disruption effect, and recovery effect). No other significant effects were found (see **Table 5**).

Stem Completion Task

We tested whether children's accuracy scores exceeded chance level (intercept at grand average across predictor levels significantly different from 0/50% probability) and whether their performance was influenced by *Targetness* and/or *Familiarity*, and whether there was an interaction with *Age*. Overall, children chose the correct stem with an accuracy of 54.94%, with individual accuracy scores ranging from 18.75 to 93.75%. **Figure 3A** shows children's individual accuracy scores along with the overall mean accuracy score for the SC task. **Figures 3B,C** show the scores according to *Familiarity* and *Targetness*, respectively. The complete output of the model is provided in **Table 6**. The corresponding estimates show that children scored significantly above chance level [intercept: log-odds = 0.22, $z = +2.17$; $p = 0.03$; 95% CI_{prob} (50.5, 60.3%)]. There were no significant differences between trials with familiar verb stems and trials with novel verb stems [*Familiarity*, log-odds = -0.15, $z = -1.02$; $p = 0.30$; 95% CI (-0.45, 0.14)] nor between target and non-target trials [*Targetness*, log-odds = 0.26, $z = 1.43$; $p = 0.15$; 95% CI (-0.11, 0.63)]. Therefore, we cannot conclude that children treat familiar items differently from novel ones or targets differently from non-targets. Moreover, there was no main effect of *Age* nor did *Age* significantly modulate the effects. No other interactions in the model yielded statistically

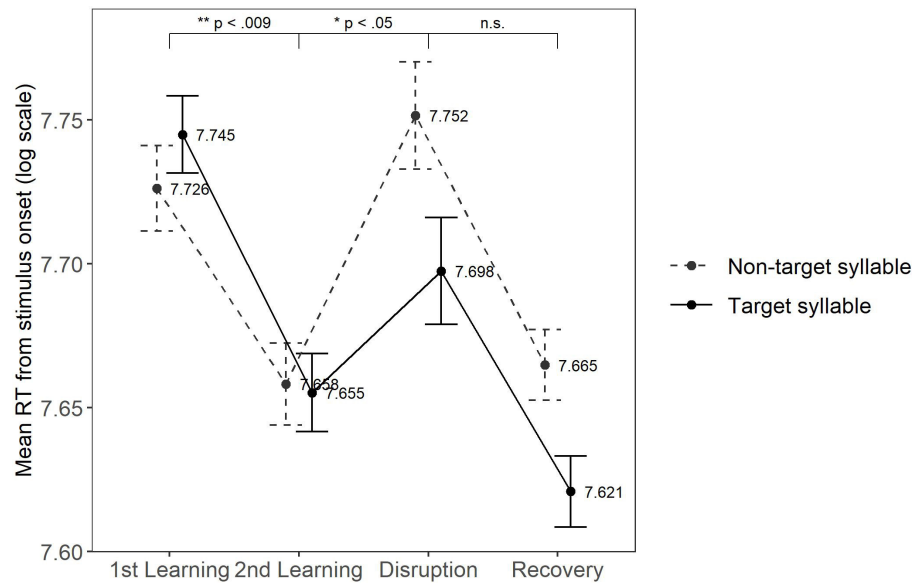


FIGURE 2 | Mean response times from stimulus onset (RTs in log-scale) and error bars with SEs for the target (solid) and non-target (dashed) syllable across the four blocks of exposure. The numerical values for the means are annotated. *Indicates statistical significance, as indicated by the p -value.

TABLE 5 | Summary of the RT model (6335 observations; $N = 52$).

| RT model – log RT | | | | | |
|---|------------|---|------------------|---|------------------|
| Predictors | Estimates* | SE* | CI* | t-Statistic | p-Value |
| (Intercept) | 7.66 | 0.04 | (7.59 to 7.74) | 195.18 | <0.001 |
| Training effect | −0.08 | 0.03 | (−0.13 to −0.02) | −2.74 | 0.006 |
| Disruption effect | 0.05 | 0.02 | (0.00 to 0.09) | 2.04 | 0.041 |
| Recovery effect | −0.04 | 0.02 | (−0.09 to 0.00) | −1.75 | 0.080 |
| Targetness | −0.02 | 0.01 | (−0.04 to −0.00) | −2.31 | 0.021 |
| Age (centered) | −0.09 | 0.03 | (−0.16 to −0.02) | −2.66 | 0.008 |
| Training effect*Targetness | −0.01 | 0.02 | (−0.06 to 0.04) | −0.44 | 0.662 |
| Disruption effect*Targetness | −0.03 | 0.04 | (−0.11 to 0.06) | −0.61 | 0.541 |
| Recovery effect*Targetness | −0.01 | 0.03 | (−0.08 to 0.06) | −0.30 | 0.765 |
| Training effect*Age (centered) | 0.04 | 0.02 | (−0.01 to 0.09) | 1.59 | 0.111 |
| Disruption effect*Age (centered) | 0.03 | 0.02 | (−0.01 to 0.07) | 1.46 | 0.143 |
| Recovery effect*Age (centered) | −0.02 | 0.02 | (−0.06 to 0.02) | −1.05 | 0.293 |
| Targetness*Age (centered) | −0.02 | 0.01 | (−0.03 to −0.00) | −2.02 | 0.043 |
| Training effect*Targetness*Age (centered) | −0.00 | 0.02 | (−0.05 to 0.04) | −0.24 | 0.813 |
| Disruption effect*Targetness*Age (centered) | 0.03 | 0.04 | (−0.04 to 0.11) | 0.87 | 0.384 |
| Recovery effect*Targetness*Age (centered) | 0.01 | 0.03 | (−0.05 to 0.07) | 0.24 | 0.808 |
| Random effects | | | | | |
| σ^2 | 0.08 | τ_1 Item.c_age | 0.00 | ρ_{01} Item | −0.91 |
| τ_0 Item | 0.00 | τ_1 Subj.trainingEffect | 0.04 | ρ_{01} Subj.trainingEffect | 0.35 |
| τ_0 Subj | 0.08 | τ_1 Subj.disruptionEffect | 0.02 | ρ_{01} Subj.disruptionEffect | 0.05 |
| N Item | 96 | τ_1 Subj.krecoveryEffect | 0.02 | ρ_{01} Subj.krecoveryEffect | −0.41 |
| N Subj | 52 | τ_1 Subj.Targetness | 0.00 | ρ_{01} Subj.Targetness | −0.55 |
| Observations: | 6335 | τ_1 Subj.trainingEffect:Targetness | 0.01 | ρ_{01} Subj.trainingEffect:Targetn | 0.26 |
| Marginal R^2 | 0.152 | τ_1 Subj.disruptionEffect:Targetness | 0.07 | ρ_{01} Subj.disr.Effect:Targetn | −0.39 |
| Conditional R^2 : | NA | τ_1 Subj.recoveryEffect:Targetness | 0.03 | ρ_{01} Subj.recoveryEffect:Targetn | 0.63 |

*All values are log-scaled. The bold values indicate that the effect was statistically significant.

significant effects (see **Table 6**). Finally, we used the method proposed by Lammertink et al. (2019, 2020) to explore whether children's performance in the SRT task correlated with their performance in the SC task. The Pearson's correlation coefficient was not statistically significantly different from zero ($r = 0.11$, $p = 0.45$).

DISCUSSION

In the present study, we examined children's ability to learn NADs within an active word-monitoring SRT task set up as a web-based computer game (*whack-a-mole*) with the objective of measuring NAD sensitivity in novel natural language stimuli. In short, our findings suggest that children between 4 and 8 years of age were sensitive to the internal rule-based structure of the two presented NADs and showed learning in both the SRT task as measured *via* RTs as well as learning in the SC task as measured *via* response accuracy.

Our findings indicate that at the group level, children in our study were able to learn the internal rule structure of both the target as well as the non-target NAD stimuli. Successful learning in the SRT word-monitoring task was indicated by: (1) a decrease in RTs during the first two learning blocks (**training effect**), suggesting that over time the correct (word-monitoring) responses to the second element of the NAD were predicted by the first element of the NAD, and by (2) an increase in RTs during the disruption block (**disruption effect**), in which violated NADs were presented (i.e., the prediction of the second element of the NAD was unreliable and thus led to increased RTs). Hence, we replicated the results of Lammertink et al. (2019, 2020) in a web-based setting. As expected, children showed overall faster responses for the target trials than the non-target trials, suggesting that children have learned the NAD related to the target syllable better compared to the non-target one. This may be due to the explicit wording of the instructions ("If you hear *are*, press the white button. If you don't hear *are*, press the red button"). Finally, our data did not show the expected **recovery effect** (i.e., RTs did not go back to baseline after the disruption block). The observation of this effect would have been a further indicator for NAD learning. This can be due to a possible lack of power caused by too much variance in the data. Alternatively, some possible reasons for the absence of this effect are that children may need more trials to "recover" from the disruption block or that they were tired and/or less attentive toward the end of the experiment, or that they were surprised in the disruption block, and therefore their recovery was weakened or hindered.

Children in the present study were also asked to complete a SC task. Success in this task required children to apply or transfer the internal rule structure of the NAD, which was learned during the SRT task, to actively access the missing second element of the NAD to the given first element and verb stem. Our results show that children's performance at the group level differed significantly from chance level, independently of targetness or familiarity of the stimuli. Children chose the correct second element of the NAD (*are* or *ando*) significantly more often than would be expected if they were only guessing. While children's

responses indicated at the group level that learning was achieved, we found large differences at the individual level, which were unrelated to age. Our SC task was a modification of the 2AFC task in Lammertink et al. (2019). In their 2AFC task participants had to decide which of the two presented utterances was more familiar to the previously heard utterances, and they failed to do so. 2AFC tasks of this kind have a high working memory load and can lead to *response biases* (Fritzley and Lee, 2003), such that children tend to provide only one type of answer (Okanda and Itakura, 2010). Hence, the SC task as used in our study might be a more suitable task for showing explicit NAD sensitivity in children, despite the high variability in the data. Finally, we have no evidence that children's SC accuracy scores could explain the variance in their SRT data because the correlation between the learning in the two tasks did not reach significance. The reason for this might be that the two tasks measure learning in different ways (implicit learning vs. accessing the knowledge more explicitly), therefore relating both might not address the same information. Also, it should be acknowledged that the lower number of trials in the SC task might distort potential effects, and therefore the SC results need to be interpreted with caution.

An additional goal of the present work was to assess whether children's NAD learning is affected by age. We found no evidence that age modulated the training effect or the disruption effect in the age range tested in our study. However, we found a main effect of age, which indicated that younger children, as expected, generally responded slower in the SRT task than older children. In addition, there was a significant interaction between age and targetness (target and non-target). That is, the youngest children (4-year-olds) were slower in responding to target items than to non-target items, while the opposite was determined for the oldest children from the sample. Furthermore, the RT distribution of the youngest children was more spread out and thus more data points from these children were excluded than from older ones. However, it is likely that this would also be the case in laboratory studies. Therefore, conclusions on the influence of age on RTs need to be considered with special caution. Our results are in line with Lammertink et al. (2019, 2020), who also showed that older children can learn NADs in this behavioral setting. In contrast to our hypothesis, we did not observe any sign of decreased sensitivity to the NADs for older children compared to younger ones, as could have been expected from the suggested developmental shift caused by the maturation of the PFC (Skeide and Friederici, 2016). A possible explanation could be that the SRT task can be considered an active task, although the NADs are not explicitly mentioned in the instruction. It might also be that case that older children coped better with the attentional and motoric demands of the web-based setting and therefore compensated for a potential age effect. We can conclude from our data that, at the group level, children between 4 and 8 years of age can learn NADs if they are assigned an active task like the one in this study (monitoring a word). We believe that future studies should explore possible underlying cognitive processes by means of brain-structural or -functional indices and should continue to address implicit vs. explicit learning in older age ranges, for example, by adding instructions pointing to the internal structure of the stimuli.

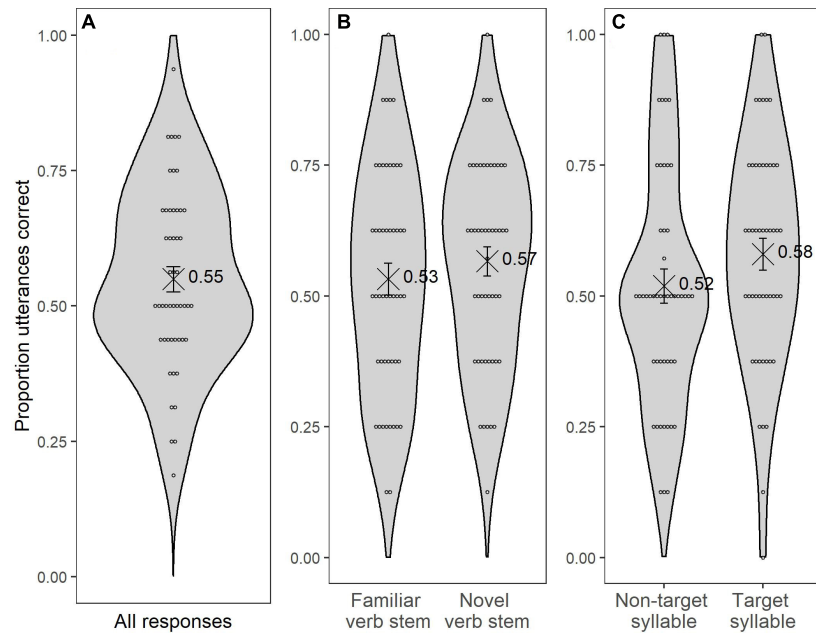


FIGURE 3 | Violin plots that represent the distribution of (A) the overall mean accuracy scores on the SC task, (B) the mean accuracy scores by *Familiarity*, and (C) the mean accuracy scores by *Targetness*. Error bars indicate SEs. The dots represent the individual scores and the black cross indicates the mean with its numerical value.

TABLE 6 | Summary of the accuracy model for the SC task (799 observations, $N = 50$).

| Accuracy | | | | | | |
|--------------------------------|---|-----------------|-------------|-----------------|-------------|--------------|
| Predictors | Log-odds | CI(log-odds) | Odds ratios | CI(odds ratios) | z-Statistic | p-Value |
| (Intercept) | 0.22 | (0.02 to 0.42) | 1.25 | (1.02 to 1.52) | 2.17 | 0.030 |
| Targetness | 0.26 | (−0.11 to 0.63) | 1.30 | (0.90 to 1.88) | 1.43 | 0.153 |
| Age (centered) | 0.11 | (−0.07 to 0.29) | 1.12 | (0.93 to 1.34) | 1.24 | 0.215 |
| Familiarity | −0.15 | (−0.45 to 0.14) | 0.86 | (0.64 to 1.15) | −1.02 | 0.306 |
| Targetness*Age (centered) | 0.15 | (−0.17 to 0.48) | 1.16 | (0.84 to 1.62) | 0.93 | 0.352 |
| Familiarity*age (centered) | −0.13 | (−0.38 to 0.13) | 0.88 | (0.68 to 1.14) | −0.96 | 0.335 |
| Random effects | | | | | | |
| σ^2 | 3.29 | | | | | |
| τ_0 Subj | 0.22 | | | | | |
| τ_1 Subj, Targetness | 0.57 | | | | | |
| ρ_{01} Subj, Targetness 1 | −0.24 | | | | | |
| ICC | 0.10 | | | | | |
| N_{Subj} | 50 | | | | | |
| Observations: 799 | Marginal R^2 /conditional R^2 : 0.015/0.112 | | | | | |

*All values are log-scaled. The bold values indicate that the effect was statistically significant.

To our knowledge, our study is the first testing children with an SRT task in a web-based setting. Here, we did not observe any substantial differences between our results from a web-based study and a similar laboratory-based experiment, as in Lammertink et al. (2019, 2020). Our study therefore demonstrates that RTs collected *via* the web with children aged 4–8 years is feasible and delivers reliable results. While running this study fully online had the advantage of allowing us to collect data during the global pandemic and in a faster manner compared

to the laboratory-based sample collection, this procedure still poses particular methodological challenges. Firstly, there are several influencing factors that cannot be controlled in the same way as in the laboratory. For example, parents were required to prepare the keyboard, assure the appropriate surroundings for their children (quiet room without distractions, appropriate volume, etc.) and were asked not to help or assist in the completion of the tasks. We presume that these prerequisites were met, but we cannot verify this. Secondly, we encountered

very long RTs, especially in younger children. The three youngest participants of our sample (aged 3;7 to 4;18) showed RTs longer than 7000 ms after stimulus offset in more than 12% of their responses. We believe that such long RTs do not reflect linguistic stimulus processing and we therefore applied different outlier criteria than typically used in a laboratory-based study (e.g., set an upper cut-off to 7000 ms and excluded values at 2.5 SD from the mean). One possibility to avoid these challenges of web-based experiments in further studies would be conducting moderated studies including a debriefing with the experimenter (at the cost of privacy data protection) or to include a time-out for the single trials. However, we found no evidence that young children in our sample had a different learning behavior than the older children and thus we assume that their large range of RTs is most likely due to their shorter attention span compared to older children and not indicative of deficits in NAD learning. Furthermore, we had to exclude a substantial number of datasets from our analyses, because children did not follow the task instruction (and pressed randomly in the word-monitoring SRT task) – a behavior that can be better monitored in a laboratory-based experiment. In future studies, those children could receive an additional practice block or instructions, or the experiment could be stopped, for instance, after the first learning block. Secondly, accuracy of RTs in a web-based setting may sometimes be considered unprecise and unreliable (Germine et al., 2012; Reimers and Stewart, 2015). Although we cannot fully control the RT precision of the *LabVanced* software, we did not notice any missing values or larger gaps in the distribution due to the software. In addition, *LabVanced* developers attest that the collection and measurement of RTs are highly reliable. Importantly, the findings from our study rely on relative RTs and not absolute RTs, which makes the measurement more reliable. We therefore encourage further studies exploring small effect sizes with RTs to consider within-subject designs. Also, to confirm the reliability of the observed results, it would be beneficial to repeat the study in a laboratory setting with a similar age sample. In this context, the 4-year-old children are of special interest, because their RT data showed the largest range. In a laboratory setting, it would be possible to evaluate some of the reasons for this behavior regarding distractibility or difficulty in focusing over a longer period of time.

In conclusion, our study provides novel evidence on the learning of NADs in children aged 4–8 years in a web-based game-like task. Despite the variability between and within participants, our data suggest that the children are capable of learning NADs in an SRT task. In addition, the present work contributes evidence to web-based research demonstrating the feasibility of testing children online. Taking the discussed advantages and challenges into consideration, we believe that the

use of online studies is a promising alternative or supplement to traditional laboratory-based studies.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://osf.io/3u62p/>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the University of Potsdam, Faculty of Humanities (EA 43/2018). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MM and AH were responsible for implementing the experimental paradigm, collecting and analyzing the data, and writing the manuscript. MM, AH, and JV were involved in data collection and data analysis. AF, CM, BH, and IW conceptualized and designed the study, discussed the implementation, and contributed to writing and revising the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the Research Unit FOR 2253: crossing the borders, Project 1 (WA2969/6-2; HO1960/18-2; FR 519/20-2; MU 3112/5-2; Project number 258522519).

ACKNOWLEDGMENTS

We would like to thank Jan Ries and Tom Fritzsche for their assistance in running and setting up the experiment and recruitment. Special thanks to Anne van der Kant for her thoughtful comments and support. We also thank the BabyLab Team and all the families that participated in the study. We gratefully acknowledge the support of the of the Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of the University of Potsdam.

REFERENCES

- Akaike, H. (1998). "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*, eds E. Parzen, K. Tanabe, and G. Kitagawa (New York, NY: Springer), 199–213. doi: 10.1007/978-1-4612-1694-0_15
- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (2020). Gorilla in our midst: an online behavioral experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x
- Arnon, I. (2019). Statistical learning, implicit learning, and first language acquisition: a critical evaluation of two developmental predictions. *Top. Cogn. Sci.* 11, 504–519. doi: 10.1111/tops.12428

- Baayen, R., and Milin, P. (2010). Analyzing reaction times. *Int. J. Psychol. Res.* 3, 12–28. doi: 10.21500/20112084.807
- Bambha, V. P., and Casasola, M. (2021). From lab to zoom: adapting training study methodologies to remote conditions. *Front. Psychol.* 12:694728. doi: 10.3389/fpsyg.2021.694728
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bialystok, E. (1986). Factors in the growth of linguistic awareness. *Child Dev.* 57:498. doi: 10.2307/1130604
- Boersma, P., and Weenink, D. (2018). *Praat: Doing Phonetics by Computer (Version 6.0.37)*. Available online at: <http://www.praat.org/> [Accessed 14 March 2018].
- Box, G., and Cox, D. (1964). An analysis of transformations. *J. R. Stat. Soc. Series B* 26, 211–252.
- Brand, A., and Bradley, A. (2012). Assessing the effects of technical variance on the statistical outcomes of web experiments measuring response times. *Soc. Sci. Comput. Rev.* 30, 350–357. doi: 10.1177/0894439311415604
- Bridges, D., Pitiot, A., MacAskill, M. R., and Peirce, J. W. (2020). The timing mega-study: comparing a range of experiment generators, both lab-based and online. *Brain Cogn.* 8:e9414. doi: 10.7717/peerj.9414
- Citron, F. M., Oberecker, R., Friederici, A. D., and Mueller, J. L. (2011). Mass counts: ERP correlates of non-adjacent dependency learning under different exposure conditions. *Neurosci. Lett.* 487, 282–286. doi: 10.1016/j.neulet.2010.10.038
- Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS One* 8:e57410. doi: 10.1371/journal.pone.0057410
- Dahan, A., Bennett, R., and Reiner, M. (2019). How long is too long: an individual time-window for motor planning. *Front. Hum. Neurosci.* 13:238. doi: 10.3389/fnhum.2019.00238
- de Leeuw, J. R., and Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and psychophysics toolbox in a visual search task. *Behav. Res. Methods* 48, 1–12. doi: 10.3758/s13428-015-0567-2
- Erickson, L. C., and Thiessen, E. D. (2015). Statistical learning of language: theory, validity, and predictions of a statistical learning account of language acquisition. *Dev. Rev.* 37, 66–108. doi: 10.1016/j.dr.2015.05.002
- Finger, H., Goeke, C., Diekamp, D., Standvoß, K., and König, P. (2017). “LabVanced: a unified JavaScript framework for online studies,” in *Proceeding of the 2017 International Conference on Computational Social Science IC2S2*, Cologne.
- Friederici, A. D., Mueller, J. L., and Oberecker, R. (2011). Precursors to natural grammar learning: preliminary evidence from 4-month-old infants. *PLoS One* 6:e17920. doi: 10.1371/journal.pone.0017920
- Friederici, A. D., Mueller, J. L., Sehm, B., and Ragert, P. (2013). Language learning without control: the role of the PFC. *J. Cogn. Neurosci.* 25, 814–821. doi: 10.1162/jocn_a_00350
- Fritzley, V. H., and Lee, K. (2003). Do young children always say yes to yes-no questions? A metadevelopmental study of the affirmation bias. *Child Dev.* 74, 1297–1313. doi: 10.1111/1467-8624.00608
- Frost, R. L., and Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition* 147, 70–74. doi: 10.1016/j.cognition.2015.11.010
- Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychon. Bull. Rev.* 19, 847–857.
- Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychol. Sci.* 13, 431–436. doi: 10.1111/1467-9280.00476
- Gómez, R. L., and Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy* 7, 183–206. doi: 10.1207/s15327078in0702_4
- Hilbig, B. E. (2016). Reaction time effects in lab- versus Web-based research: experimental evidence. *Behav. Res. Methods* 48, 1718–1724. doi: 10.3758/s13428-015-0678-9
- Höhle, B., Schmitz, M., Santelmann, L. M., and Weissenborn, J. (2006). The recognition of discontinuous verbal dependencies by German 19-month-olds: evidence for lexical and structural influences on children's early processing capacities. *Lang. Learn. Dev.* 2, 277–300.
- Isbilen, E. S., McCauley, S. M., Kidd, E., and Christiansen, M. H. (2017). “Testing statistical learning implicitly: A novel chunk-based measure of statistical learning,” in *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. (Austin, TX: Cognitive Science Society).
- Kidd, E., and Arculi, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Dev.* 87, 184–193.
- Kliegl, R., Masson, J., and Richter, E. M. (2009). A linear mixed model analysis of masked repetition priming. *Vis. Cogn.* 18, 655–681. doi: 10.1080/13506280902986058
- Kochari, A. R. (2019). Conducting web-based experiments for numerical cognition research. *J. Cogn.* 2:39. doi: 10.5334/joc.85
- Kowarik, A. (2019). *boxcox: Boxcox Power Transformation*. Available online at: <https://www.rdocumentation.org/packages/EnvStats/versions/2.3.1/topics/boxcox#:~:text=Boxcox%20Power%20Transformation,based%20on%20a%20specified%20objective> (accessed August 15, 2021).
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13
- Lammertink, I., Boersma, P., Wijnen, F., and Rispens, J. (2020). Children with developmental language disorder have an auditory verbal statistical learning deficit: evidence from an online measure. *Lang. Learn.* 70, 137–178. doi: 10.1111/lang.12373
- Lammertink, I., van Witteloostuijn, M., Boersma, P., Wijnen, F., and Rispens, J. (2019). Auditory statistical learning in children: novel insights from an online measure. *Appl. Psycholinguist.* 40, 279–302. doi: 10.1017/S0142716418000577
- Lany, J., and Shoaib, A. (2020). Individual differences in non-adjacent statistical dependency learning in infants. *J. Child Lang.* 47, 483–507. doi: 10.1017/S0305000919000230
- López-Barroso, D., Cucurell, D., Rodríguez-Fornells, A., and de Diego-Balaguer, R. (2016). Attentional effects on rule extraction and consolidation from speech. *Cognition* 152, 61–69. doi: 10.1016/j.cognition.2016.03.016
- Marchetto, E., and Bonatti, L. L. (2013). Words and possible words in early language acquisition. *Cogn. Psychol.* 67, 130–150. doi: 10.1016/j.cogpsych.2013.08.001
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing Type I error and power in linear mixed models. *J. Mem. Lang.* 94, 305–315. doi: 10.1016/j.jml.2017.01.001
- Morini, G., and Blair, M. (2021). Webcams, songs, and vocabulary learning: a comparison of in-person and remote data collection as a way of moving forward with child-language research. *Front. Psychol.* 12:702819. doi: 10.3389/fpsyg.2021.702819
- Mueller, J. L., Friederici, A. D., and Männel, C. (2012). Auditory perception at the root of language learning. *Proc. Natl. Acad. Sci. U.S.A.* 109, 15953–15958. doi: 10.1073/pnas.1204319109
- Mueller, J. L., Friederici, A. D., and Männel, C. (2019). Developmental changes in automatic rule-learning mechanisms across early childhood. *Dev. Sci.* 22:e12700. doi: 10.1111/desc.12700
- Mueller, J. L., Milne, A., and Männel, C. (2018). Non-adjacent auditory sequence learning across development and primate species. *Curr. Opin. Behav. Sci.* 21, 112–119. doi: 10.1016/J.COBEHA.2018.04.002
- Mueller, J. L., Oberecker, R., and Friederici, A. D. (2009). Syntactic learning by mere exposure—an ERP study in adult learners. *BMC Neurosci.* 10:89. doi: 10.1186/1471-2202-10-89
- Nissen, M. J., and Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. *Cogn. Psychol.* 19, 1–32. doi: 10.1016/0010-0285(87)90002-8
- Nussenbaum, K., Scheuplein, M., Phaneuf, C., Evans, M., and Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra* 6:17213. doi: 10.1525/collabra.17213
- Okanda, M., and Itakura, S. (2010). When do children exhibit a “yes” bias? *Child Dev.* 81, 568–580. doi: 10.1111/j.1467-8624.2009.01416.x
- Paul, M., Männel, C., van der Kant, A., Mueller, J. L., Höhle, B., Wartenburger, I., et al. (2020). Gradual development of non-adjacent dependency learning during early childhood. *bioRxiv [Preprint]* doi: 10.1101/2020.09.01.277822
- Qian, T., Jaeger, T. F., and Aslin, R. N. (2016). Incremental implicit learning of bundles of statistical patterns. *Cognition* 157, 156–173. doi: 10.1016/j.cognition.2016.09.002

- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Core Team.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychol. Bull.* 114, 510–532. doi: 10.1037/0033-2909.114.3.510
- Reimers, S., and Stewart, N. (2015). Presentation and response timing accuracy in adobe flash and HTML5/JavaScript Web experiments. *Behav. Res. Methods* 47, 309–327. doi: 10.3758/s13428-014-0471-1
- Rhodes, M., Rizzo, M., Foster-Hanson, E., Moty, K., Leshin, A. R., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21:4. doi: 10.1080/15248372.2020.1797751
- Santelmann, L. M., and Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: evidence for limitations in processing space. *Cognition* 69, 105–134. doi: 10.1016/s0010-0277(98)00060-2
- Sauter, M., Draschkow, D., and Mack, W. (2020). Building, hosting and recruiting: a brief introduction to running behavioral experiments online. *Brain Sci.* 10:251. doi: 10.3390/brainsci10040251
- Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. *J. Mem. Lang.* 110:104038. doi: 10.1016/j.jml.2019.104038
- Scott, K., Chu, J., and Schultz, L. (2017). Lookit (Part 2): assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/OPMI_a_00001
- Silver, A., Elliott, L., Braham, E., Bachman, H., Votruba-Drzal, E., Tamis-LeMonda, C., et al. (2021). Measuring emerging number knowledge in toddlers. *Front. Psychol.* 12:703598. doi: 10.3389/fpsyg.2021.703598
- Simcox, T., and Fiez, J. A. (2014). Collecting response times using Amazon mechanical Turk and adobe flash. *Behav. Res. Methods* 46, 95–111. doi: 10.3758/s13428-013-0345-y
- Skeide, M. A., and Friederici, A. D. (2016). The ontogeny of the cortical language network. *Nat. Rev. Neurosci.* 17, 323–332. doi: 10.1038/nrn.2016.23
- Uddén, J., Ingvar, M., Hagoort, P., and Petersson, K. M. (2012). Implicit acquisition of grammars with crossed and nested non-adjacent dependencies: investigating the push-down stack model. *Cogn. Sci.* 36, 1078–1101. doi: 10.1111/j.1551-6709.2012.01235.x
- Vales, C., Wu, C., Torrance, J., Shannon, H., States, S., and Fisher, A. (2021). Research at a distance: replicating semantic differentiation effects using remote data collection with children participants. *Front. Psychol.* 12:697550. doi: 10.3389/fpsyg.2021.697550
- van der Kant, A., Männel, C., Paul, M., Friederici, A. D., Höhle, B., and Wartenburger, I. (2020). Linguistic and non-linguistic non-adjacent dependency learning in early development. *Dev. Cogn. Neurosci.* 45:100819. doi: 10.1016/j.dcn.2020.100819
- van Heugten, M., and Shi, R. (2010). Infants' sensitivity to non-adjacent dependencies across phonological phrase boundaries. *J. Acoust. Soc. Am.* 128, EL223–EL228. doi: 10.1121/1.3486197
- Wang, F. H., Zevin, J., and Mintz, T. H. (2019). Successfully learning non-adjacent dependencies in a continuous artificial language stream. *Cogn. Psychol.* 113:101223. doi: 10.1016/j.cogpsych.2019.101223
- Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis. Use R!*. Berlin: Springer.
- Wilson, B., Spierings, M., Ravnani, A., Mueller, J. L., Mintz, T. H., Wijnen, F., et al. (2018). Non-adjacent dependency learning in humans and other animals. *Top. Cogn. Sci.* 12, 843–858. doi: 10.1111/tops.12381
- Yang, C. (2004). Universal grammar, statistics, or both. *Trends Cogn. Sci.* 8, 451–456.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Marimon, Hofmann, Verissimo, Männel, Friederici, Höhle and Wartenburger. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Moderated Online Data-Collection for Developmental Research: Methods and Replications

Aaron Chuey^{1*}, Mika Asaba¹, Sophie Bridgers², Brandon Carrillo¹, Griffin Dietz¹, Teresa Garcia¹, Julia A. Leonard³, Shari Liu², Megan Merrick⁴, Samaher Radwan¹, Jessa Stegall¹, Natalia Velez⁵, Brandon Woo⁵, Yang Wu¹, Xi J. Zhou¹, Michael C. Frank¹ and Hyowon Gweon^{1*}

¹ Department of Psychology, Stanford University, Palo Alto, CA, United States, ² Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States, ³ Department of Psychology, Yale University, New Haven, CT, United States, ⁴ Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, United States, ⁵ Department of Psychology, Harvard University, Cambridge, MA, United States

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Lisa Feigenson,
Johns Hopkins University,
United States
Norbert Zmyj,
Technical University Dortmund,
Germany

*Correspondence:

Aaron Chuey
chuey@stanford.edu
Hyowon Gweon
gweon@stanford.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 01 July 2021

Accepted: 06 October 2021

Published: 03 November 2021

Citation:

Chuey A, Asaba M, Bridgers S,
Carrillo B, Dietz G, Garcia T,
Leonard JA, Liu S, Merrick M,
Radwan S, Stegall J, Velez N, Woo B,
Wu Y, Zhou XJ, Frank MC and
Gweon H (2021) Moderated Online
Data-Collection for Developmental
Research: Methods and Replications.
Front. Psychol. 12:734398.
doi: 10.3389/fpsyg.2021.734398

Online data collection methods are expanding the ease and access of developmental research for researchers and participants alike. While its popularity among developmental scientists has soared during the COVID-19 pandemic, its potential goes beyond just a means for safe, socially distanced data collection. In particular, advances in video conferencing software has enabled researchers to engage in face-to-face interactions with participants from nearly any location at any time. Due to the novelty of these methods, however, many researchers still remain uncertain about the differences in available approaches as well as the validity of online methods more broadly. In this article, we aim to address both issues with a focus on moderated (synchronous) data collected using video-conferencing software (e.g., Zoom). First, we review existing approaches for designing and executing moderated online studies with young children. We also present concrete examples of studies that implemented choice and verbal measures (Studies 1 and 2) and looking time (Studies 3 and 4) across both in-person and online moderated data collection methods. Direct comparison of the two methods within each study as well as a meta-analysis of all studies suggest that the results from the two methods are comparable, providing empirical support for the validity of moderated online data collection. Finally, we discuss current limitations of online data collection and possible solutions, as well as its potential to increase the accessibility, diversity, and replicability of developmental science.

Keywords: online research, cognitive development, meta-analysis, replication, moderated data collection

INTRODUCTION

Over the past decade, online data collection has transformed the field of psychological science. Commercial crowdsourcing platforms such as Amazon Mechanical Turk have allowed participants to perform experimental tasks remotely from their own computers, making it easier, faster, and cheaper for researchers to collect large samples. The advantages of online methods led to a rapid increase in their popularity; for example, the percentage of online studies published in three

prominent social psychology journals rose from around 3% in 2005 to around 50% in 2015 (Anderson et al., 2019).

Although online methods have been mostly constrained to studies with adults, some recent efforts have pioneered ways to conduct developmental research online (e.g., Lookit, Scott and Schulz, 2017; TheChildLab.com, Sheskin and Keil, 2018; Panda, Rhodes et al., 2020). As the COVID-19 pandemic spurred many developmental researchers to consider safer alternatives to in-person interactions, these methods have quickly gained traction as an innovative way to enable large-scale data collection from children and maximize access and impact in developmental science (Sheskin et al., 2020). Due to the novelty of these methods, however, there is little shared information available about recommended practices for designing, implementing, and executing online experiments with children. Furthermore, researchers may feel hesitant to replicate or build on prior work using online methods because of uncertainties about how developmental data collected online would compare to data collected in person.

The current paper aims to serve as a guide for developmental researchers seeking information about online data collection, with a focus on using video-chat software for moderated (synchronous) data collection. We begin by explaining how moderated methods differ from unmoderated (asynchronous) methods, including their relative advantages and disadvantages. Next, we describe recommended practices and approaches for designing online developmental studies conducted via moderated sessions. In particular, we provide guidelines for implementing two broad classes of measures: forced choice for young children and looking time for infants. To examine the validity of moderated online methods, we present four sets of studies conducted both in person and online that utilize these measures as well as a meta-analysis that compares results from both data collection methods across the four sets of studies. Finally, we discuss the limitations and potential of moderated online data collection as a viable research method that will continue to shape developmental psychology.

MODERATED ONLINE STUDIES: WHAT IT IS AND RECOMMENDED PRACTICES

Online data collection methods can be categorized as moderated (synchronous) or unmoderated (asynchronous). Unlike **unmoderated (asynchronous) data collection** which functions like Amazon Mechanical Turk, **moderated (synchronous) data collection** functions more like in-person testing; participants engage in real-time interactions with researchers on a web-enabled device using video-conferencing software, such as Zoom, Adobe Connect, or Skype.

An advantage of unmoderated data collection is that it is less labor-intensive than moderated data collection. Participants complete a preprogrammed module without directly interacting with researchers; once the study is programmed, there is little effort involved in the actual data collection process on the researchers' end. Some pioneering efforts have led to innovative platforms for implementing these modules (Lookit,

see Scott and Schulz, 2017; see also Panda, Rhodes et al., 2020), and adaptations of three well-established studies on Lookit have found comparable results to their original in-person implementations (Scott et al., 2017). Its advantages, however, come with trade-offs: due to the lack of researcher supervision, unmoderated data collection is limited to behavioral paradigms where real-time monitoring is not necessary. Thus, this method may not be as well suited for studies where live social interactions and joint-attention are central to the hypothesis and experimental design. Furthermore, adapting an in-person study to an unmoderated module usually involves significant alterations in study procedure and format (Scott et al., 2017), creating additional challenges to directly replicating existing findings in some circumstances.

Moderated data collection, by contrast, is comparable to in-person methods in terms of their costs. It requires recruiting and scheduling participants for an appointment, and at least one researcher must be available to host the session and guide participants throughout the study procedure. Yet, this allows moderated sessions to retain the interactive nature of in-person studies that is often critical for developmental research. Experimenters can have face-to-face interactions with parents and children to provide instructions, present stimuli, actively guide children's attention, ask questions, and record a number of behavioral measures. Although certain paradigms or measures are difficult to implement even with moderated methods (e.g., playing with a physical toy), many existing in-person studies can be translated into an online version with relatively minor changes in procedures.

Early efforts to apply moderated online data-collection to studies with children have produced promising results, albeit with some caveats. For instance, Sheskin and Keil (2018) collected verbal responses from 5- to 12-year-old children in the United States on several basic tasks via video-conferencing software (Adobe Connect). While children showed ceiling-level performance on questions that assessed their understanding of basic physical principles (e.g., gravity) and fair distribution of resources, their performance on false belief scenarios (i.e., the Sally-Anne task adapted from Baron-Cohen et al., 1985) was significantly delayed compared to results from prior work conducted in person. It is possible that younger children found it more difficult to keep track of multiple characters and locations on a completely virtual interaction; the task also relied primarily on verbal prompts without additional support to guide children's attention (e.g., pointing). However, because this study did not directly compare the results from online and in-person versions of the same task, it is difficult to draw strong conclusions about the cause of the discrepancies or the validity of moderated methods more generally.

More recently, Smith-Flores et al. (2021) reported replications of prior looking time studies with infants (violation of expectation and preferential looking) via a moderated online format. The findings from data collected online were generally comparable to existing results; for instance, infants looked longer at events where an object violated the principle of gravity than events that did not (e.g., Spelke et al., 1992) and were more likely to learn about object properties following such surprising events

(Stahl and Feigenson, 2015)¹. Contrary to classic work on infants' understanding of physics, however, infants in this study did not show a sensitivity to violations of object solidity. Although infants in these experiments viewed recorded video clips of events very similar to those used in prior in-person studies, the authors note the experience of viewing such videos on screens is quite different from viewing the event in person, and that differences in the visual properties of test stimuli (e.g., limited aspect ratio of participants' screens) could have contributed to the discrepancy in results. These concerns might apply to any study using online data collection (both moderated and unmoderated) that involves viewing visual stimuli on a screen as opposed to live events.

In sum, existing data suggest that moderated online studies are indeed feasible, but they also highlight two challenges. First, due to the relative novelty of moderated methods, researchers may be unsure about how to implement a study online and what can be done to minimize potential discrepancies between in-person and online versions. Second, the field still lacks a true apples-to-apples comparison between studies conducted online and in-person using stimuli and procedures matched as closely as possible. In particular, given the variety of dependent measures and procedures used in developmental research, it is important to have a number of such comparisons that span across different experimental designs and methods.

The following sections address these challenges by reviewing current approaches to moderated online study design and providing empirical data that replicate in-person findings with moderated online methods. We begin by outlining key considerations for implementing moderated studies, followed by presentation methods and design considerations that promote participant attention and engagement. Then, we provide concrete examples of implementing dependent measures that are frequently used in developmental research: choice and verbal measures (more suitable for children aged 2 and up) and looking time measures (suitable for infants). We also compare results from experiments that were conducted in-person and adapted for online data collection using these suggestions.

MODERATED ONLINE STUDIES: IMPLEMENTATION AND RECOMMENDED PRACTICES

Moderated online studies have been implemented using a variety of video-conferencing software, including Zoom, Adobe Connect, and Skype, among others. Each video conferencing software has benefits and drawbacks that make it better suited for certain research endeavors and styles. There are several particularly important dimensions to consider, including accessibility, functionality, and robustness to technical issues (see **Table 1**).

One common way to implement moderated online studies with young children utilizes locally installed slideshow applications on experimenters' computers (e.g., Microsoft

PowerPoint, Keynote). These applications allow researchers to present a wide variety of stimuli, including images, animations, videos, audio, and written language. Implementing studies using these applications creates a linear structure that naturally segments study procedures into manageable components, making it easy for researchers to manipulate the order of presentation and access notes. Alternatively, studies involving videos, such as many infants studies, have been implemented on video-sharing websites such as YouTube, or slides hosted on cloud services.

One key challenge in designing developmental experiments is ensuring that children stay engaged and attentive throughout the task. On the one hand, an advantage of online data collection is that children participate from their familiar home environments, which could improve their comfort and engagement. On the other hand, however, home environments can be more distracting than lab settings, and researchers have little control over them. For studies that require relatively well-controlled environments, researchers could consider sending parents instructions prior to the testing session to help them create ideal testing environments at home. For example, parents could be instructed to keep siblings out of the room during the session. Here we discuss a few additional strategies to maximize children's engagement during online data collection and to direct their attention to specific stimuli on screen.

Elicit Regular Responses From Participants

Because online studies can suffer from technical problems as well as distractions in a child's home environment, researchers should design them to be robust to frequent interruptions. Eliciting regular feedback from children, either casually or by implementing comprehension questions throughout the task, is one useful strategy. While this is also used in person, frequent questions are particularly useful for identifying long periods of lag or technical issues that can otherwise go unnoticed online. Playing a short video at the start of a session and asking participants to report any lag or audio problems is another quick and easy way to assess participant-end technical issues that might not be readily apparent from an experimenter's perspective. Finally, it is often useful to make parts of a study easy to repeat in case they are compromised by connectivity issues, audio/video problems, or other unexpected difficulties.

Use Social Cues

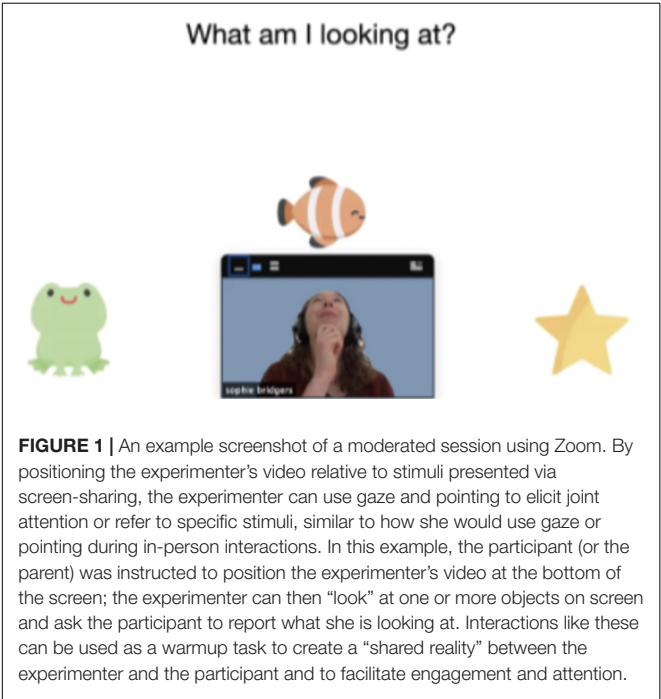
In-person studies often utilize social cues from the experimenter (e.g., gaze, pointing) to direct children's attention. While these are more difficult to use online, some video conferencing software (e.g., Zoom) allows researchers to flexibly adjust the size and location of experimenter's video feed on participants' screen, such that the experimenter's gaze and pointing can be "directed" to specific parts of the stimuli (see **Figure 1**). These features can be useful for providing the experience of a "shared reality" with the experimenter and can be particularly effective in studies that require joint attention. Additionally, audio and visual attention-getters (e.g., sounds, animations, or markers like bounded boxes

¹ This study also included a successful replication of Wu et al. (2017) on infants' understanding of emotional expressions.

TABLE 1 | Factors to consider when choosing software for moderated online data collection.

| | |
|---------------|--|
| Accessibility | Software should ideally be easy to obtain and use, especially for participants. In addition to monetary concerns or internet access (Lourenco and Tasimi, 2020), the need for technical skills, time (e.g., for downloading and installing new software), or specific hardware (e.g., Facetime requires Apple OS) can create barriers to participation. Intuitive software also makes online research easier for both experimenters and participants by reducing time spent setting up and troubleshooting sessions. Using software that many people already have and know how to use can alleviate this issue. Note, however, that accessibility is always relative to a particular population at a particular time; software that is suitable for one population may not necessarily be so for others. For example, Zoom became a more accessible option for conducting developmental research in the United States following the COVID-19 pandemic as more families downloaded and used Zoom in their day-to-day lives for work and remote schooling. As trends in software usage change over time for a given population, researchers should continue to adapt their methodologies accordingly. |
| Functionality | A software's user interface, customizability, and security features determine how studies are conducted and the extent to which researchers can customize participants' online experience. Importantly, security standards regarding recording and storage of online sessions vary across institutions and countries; researchers should keep these in mind when assessing the level of security a given software provides. Additionally, while basic video- and screen-sharing as well as text-chat functionalities are common in most software, the details vary in a number of ways, including how users customize what they can view on screen and how recording is implemented (e.g., local vs. cloud storage). More broadly, intuitive design and real-time flexibility often trades off with precise structure and customization options. Some software (e.g., Adobe Connect) allows experimenters to predetermine the layout of participants' screens before sessions, and others (e.g., Zoom) automatically generate participants' layouts and allow participants to modify their layout in real time (following instructions from experimenters). While the former type is ideal for experiments that require precise control over what participants view on screen, the latter type of software is more suitable for sessions involving rapid transitions between multiple experiments with different visual layouts. |
| Robustness | Recurring lag, audio or video problems, and even login errors can slow down or derail an online session. Although technical issues can also occur in person, issues can be more difficult to resolve in remote interactions where experimenters have limited means to understand participants' issues. Therefore, it is important to test the frequency and duration of technical issues on both experimenters' and participants' ends before committing to a particular video-conferencing software. Depending on the software, screen-sharing or streaming large video or audio files can contribute to unwanted lag or delays. Further, their severity can vary depending on connection speed or devices used by both experimenters and participants. For experiments that rely on precise timing of presented stimuli, researchers might consider presentation methods that do not rely on screen-sharing (e.g., hosting video stimuli on servers or other platforms where participants can access directly, such as online video-hosting or slide-presentation services). If there are consistent participant-end issues that impact the fidelity of a study, researchers can also set explicit criteria for participation (e.g., must use a laptop or cannot use a phone signal-based internet connection). |

that highlight a particular event, character, or object on the screen) can be used instead of experimenters' gaze or pointing gestures to focus children's attention on specific stimuli.



Keep It Short and Simple

Because interacting with others online can tax children's (and adults') cognitive resources more than in-person interactions (e.g., Bailenson, 2021), it is important to keep online studies as short and simple as possible. For studies that require relatively longer sessions, presenting them as a series of multiple, distinct activities can help maintain children's attention and enthusiasm throughout. In cases where concerns about cross-study contamination are minimal, researchers can also run more than one experiment per session. Of course, different studies have different attentional demands and require varying levels of continuous attention. Thus, researchers should consider what counts as a consequential lapse of attention and devise their exclusion criteria accordingly during the pre-registration process.

As we emphasized earlier, one key advantage of moderated methods is the relative ease of adapting in-person studies to an online format without significant changes to the procedure. This means that many of the strategies used to promote attention and engagement in person also apply to online studies. For instance, color-coding and animating stimuli, using engaging stories and characters, and talking in simple, plain language can also help children stay engaged. Overall, relatively minor changes to the way that stimuli are presented can have a large impact on children's attention and engagement throughout an online session.

In what follows, we provide more specific guidelines for implementing two kinds of dependent measures (choice and looking time) with concrete example studies for each type of

measure. Importantly, these studies address different theoretical questions and have not been fully published at the time of writing this article; the key reason for reporting these datasets is to examine the validity of moderated online data collection. As such, we describe the hypotheses and methods of these studies only to the degree necessary to contextualize our analyses: comparing the main effect of interest from data collected in-person versus online. In addition to a direct comparison of their results, we present a meta-analysis of all four sets of experiments that provides further evidence that moderated online and in-person testing yielded similar results across the current studies.

EXAMPLES AND REPLICATIONS I: CHOICE AND VERBAL MEASURES IN MODERATED ONLINE STUDIES

To elicit explicit choices from children who are old enough to understand verbal instructions, in-person studies often use pointing or reaching as dependent measures. These behaviors, however, can be difficult to assess in online studies; webcam placement can vary across participants, and participants may move outside the field of view during the critical response period. One useful approach for implementing choice tasks for children in this age range is to replace pointing or reaching with verbal responses, and associate each choice with overt visual cues such as color. For example, a binary choice question can be presented as a choice between one character wearing orange and another character wearing purple (color assignment counterbalanced), with children only needing to choose “orange or purple” (see **Figure 2**). In these choice paradigms, it is important to keep

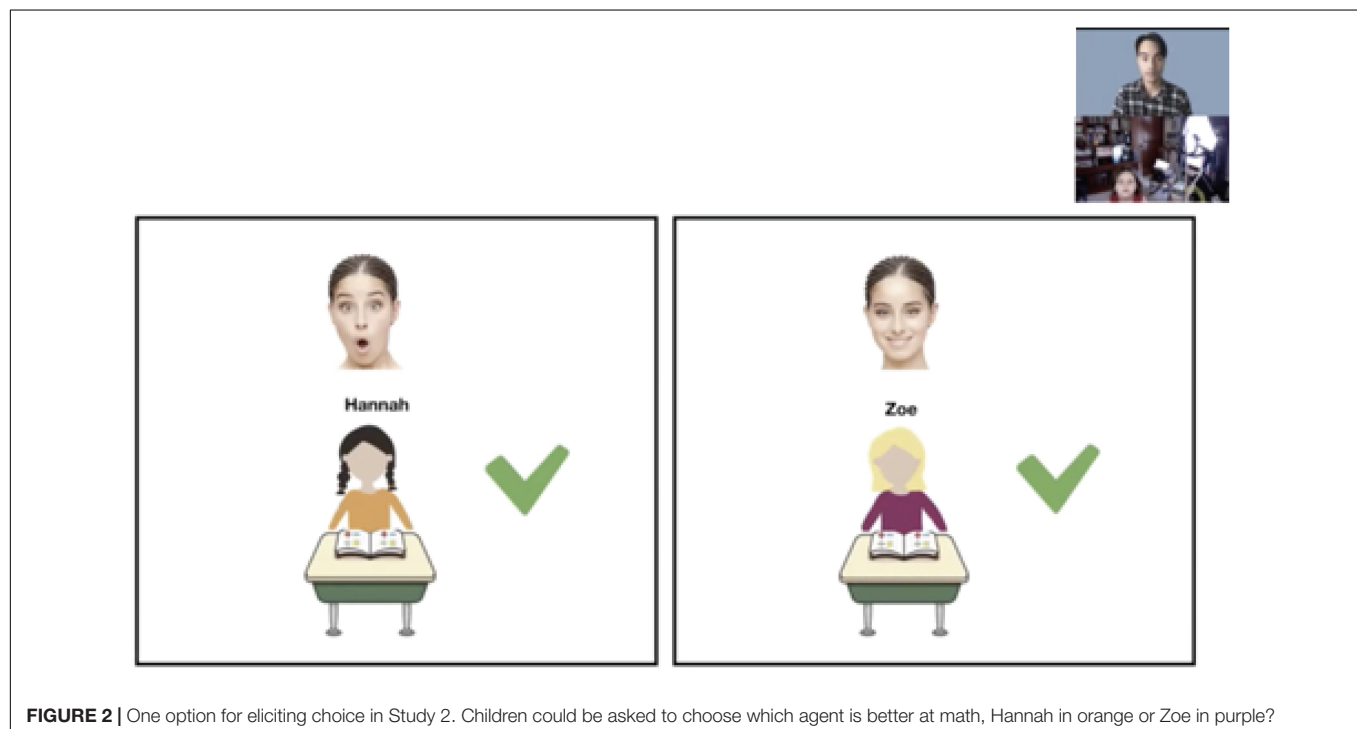
the on-screen location of key choices or stimuli as consistent as possible throughout the study such that transitioning between slides is less disruptive and easier to follow.

In addition to forced-choice measures, experimenters can elicit free-form verbal responses or actions as dependent measures, or ask the parent to type out the child’s responses via text chat. Researchers can also implement other creative dependent measures, such as prompting children to make a drawing and share it with the experimenter via video. As long as a behavior can be consistently prompted and recorded, it can likely be used as a measure in a moderated online study. Here, we present two additional sets of studies conducted online and in person that measured children’s explicit choice between two agents. One study examined 4- to 5-year olds (Study 1) and another examined 6- to 9-year olds (Study 2).

Study 1

Research Question

Can 4- to 5-year-old children use information about task difficulty to infer relative competence when agents’ efforts are matched? To investigate this question, children viewed two agents who used 10 wooden blocks to build different structures; one placed the blocks on top of each other to create a vertical tower while the other placed them next to each other to form a horizontal line. Children were then asked which agent was better at building blocks. Prior work has established that children understand that the vertical structure is “harder” (i.e., takes longer) to build compared to the horizontal structure (Gweon et al., 2017). Thus, the hypothesis was that even though both agents moved and placed 10 blocks, if they took equally long to finish building, children would judge the agent who built the



vertical (and therefore harder) tower as more competent than the agent who built the horizontal line.

Participants

In-person

Twenty 4- and 5-year-old children participated in-person at the Boston Children's Museum (10 females, mean: 62.25 months, range: 49–71); 10 additional children were tested but excluded due to failing the practice question ($n = 3$) or inclusion criteria question ($n = 7$).

Online

Twenty 4- and 5-year-old children participated online (nine females, mean: 59.23 months, range: 49–71). Participants were recruited via local and online advertising. Seven additional children were tested but excluded due to failing the inclusion criteria question ($n = 3$), technical issues ($n = 1$), declining video ($n = 1$), not wanting to answer the inclusion question ($n = 1$) or dropping out ($n = 1$).

Methods

Both in-person and online

An experimenter first asked children “Who is better at writing letters — you or your parents?” and then “Who is better at playing on the playground — you or your parents?” If children chose themselves for writing or their parents for playing, they were corrected. In the test video, children watched two agents build block structures. Below one agent was a picture of a 10-block vertical tower and below the other agent was a picture of a 10-block horizontal tower. We chose these structures based on findings from Gweon et al. (2017) showing that 4-year olds readily judge the 10-block vertical structure as harder to build than the 10-block horizontal structure based on static pictures of the initial states (i.e., scattered blocks) and final states (finished towers), without seeing the building process. The agents first said they wanted to build a pictured tower. One agent pointed to the picture below her and said, “I’m going to make this,” then the other agent repeated the same action. Next, the agents began to build at the same time. A screen blocked visual access to the agents’ building actions. Both agents indicated they were finished building at the same time. The screen then lifted, revealing what each agent made. Children were then asked the test question followed by an additional question used as a part of the inclusion criteria. Those who answered the inclusion question inaccurately were excluded from analyses.

In-person

Before the test trial, children watched a practice video where two agents drew shapes, finishing at different times. While the agents drew, a screen blocked them. One of the agents indicated she was done drawing, followed by the other agent a few seconds later. Then the screen lifted to reveal what they made. Children were asked which agent finished first and whether the agents had made the same or different pictures. If they answered incorrectly, they were excluded from analysis. Afterward, children viewed the test trial, and were subsequently asked the critical test question: “Who is better at building blocks?” and were encouraged to point, followed by the inclusion question “Which tower is better?”

Online

The online study was the same as the in-person study except for the following modifications. To make the study amenable to online testing, children’s attention toward desired locations in the presentation was cued using animation and sound. Instead of asking children to point to which agent was better at the end, they were instructed to make their choice based on the color of squares surrounding each agent. To reduce study time, the practice trial was also removed (more than 93% of children passed the practice trial in in-person versions of three similar prior studies). Finally, we changed the inclusion question to “which tower is harder to make?”

Results

In-person

Children’s performance on the test question was significantly above chance (90%, $CI = [80\%, 100\%]$, $p < 0.001$). This result held even after including the seven children who failed to answer the inclusion question accurately (74%, $CI = [60\%, 93\%]$, $p = 0.02$).

Online

Consistent with in-person findings, children’s performance on the test question was significantly above chance (85%, $CI = [70\%, 100\%]$, $p = 0.003$). See **Figure 4 (1)** for a summary of results.

Study 2

Research Question

Do children use an adult’s expressions of surprise to draw inferences about others’ competence? The in-person data was first reported in a study by Asaba et al. (2020). Children were shown two students who both succeeded or failed at a task (e.g., a math problem), accompanied by their teacher’s reaction; the teacher responded with a surprised expression to one and an unsurprised expression to the other. Children were then asked which student was better at the task. The hypothesis was that children would use the teacher’s surprise to infer the students’ competence; a teacher’s surprise at a student’s failure likely indicates competence whereas the same surprised expression in response to a student’s success indicates a lack of competence.

Participants

In-person

Twenty-eight 4- to 9-year-old children (mean = 79.2 months, range 49.2–118.8 months; 13 girls, 15 boys) participated in-person at a museum ($n = 20$) and campus preschool ($n = 8$) in Palo Alto, CA, United States. Participants who did not respond to the test questions (i.e., responded “both” to all questions; $n = 16$ -year-old) were excluded.

Online

Ninety 6- to 8-year-old children (30 6-year olds, 30 7-year olds, 30 8-year olds; mean age = 90 months, range = 72–106.8 months; 48 girls boys, 42 boys girls). Participants were recruited via local and online advertising. An additional child was tested and excluded due to having audio problems during the testing session (pre-registered exclusion).

Methods

Both in-person and online

Subjects were first introduced to a teacher and her two facial expressions, described as “surprised” and “non-surprised,” respectively. Then, all participants underwent the key trials. In each trial, two students either both succeeded or failed an activity, and the teacher expressed surprise to one student while expressing no surprise to the other (henceforth “surprise student” and “non-surprise student,” respectively). Specifically, the experimenter first remarked on one student’s performance outcome (either a success: “Look, Hannah got the math problem right!” or a failure: “Look, Hannah got the math problem wrong!”), revealed the teacher’s emotional response to the outcome (either a surprised or non-surprised face), and asked a check question: “Is the teacher surprised or not surprised?” If participants provided an incorrect response to the check question, the experimenter corrected them. This sequence was repeated for the other student (e.g., “Zoe”; gender-matched) in the trial who performed exactly the same but received the other emotional response. Finally, with images of the students’ outcomes and the teacher’s expressions visible, the experimenter asked, “One of the kids is better at this game. Who is better?” Children then indicated their response.

In-person

Children viewed eight trials, consisting of four different activities (math, spelling, kicking, and throwing) and two types of outcomes (success, failure) for each. After each trial, children indicated their response by pointing or responding verbally with the student’s name.

Online

Children only saw four trials instead of eight to reduce the length of the online experiment. The four trials consisted of two activities (randomly selected from the four activities in the in-person study) with two types of outcomes for each. Participants responded by saying the student’s name aloud.

Results

In-person

As a group, children chose the non-surprise student in success trials (71.4%, $Z = 2.91$, $p = 0.004$, Exact Wilcoxon-Pratt Signed-Rank Test), but did not choose the surprise student in fail trials significantly above chance (59.8%, $Z = 1.32$, $p = 0.211$). Given the wide age range, children were median-split into younger (age: 4.1–5.9; $N = 14$) and older age groups (age: 6.2–9.9; $N = 14$) and children’s choices within each trial were examined. The older group was accurate for both success and fail trials (Success: 98.2%, $Z = 3.64$, $p < 0.001$; Fail: 76.8%, $Z = 2.16$, $p = 0.039$). The younger group was at chance for both trial types (Success: 44.6%, $Z = -0.58$, $p = 0.71$; Fail: 42.9%, $Z = -0.81$, $p = 0.54$) with no difference between success and fail trials ($Z = 0.06$, $p = 0.99$).

Online

As a group, participants (age: 6.0–8.9; $N = 90$) chose the non-surprise student in the success trials as more competent (Mean proportion = 68.33%, $Z = 3.62$, $p < 0.001$ (Wilcoxon Signed-Rank Test) and the surprise student in fail trials (Mean

proportion = 84.44%, $Z = 6.85$, $p < 0.001$, Wilcoxon Signed-Rank Test). These results are comparable to the results of the older group (age: 6–9) in the in-person study. Note that the online version of the task found an interaction between age and outcome, whereas in-person studies additionally found significant main effects of outcome and age. However, this difference may be due to the fact that online studies had a more limited age range. See **Figure 4 (2)** for a summary of results.

EXAMPLES AND REPLICATIONS II: LOOKING TIME MEASURES IN MODERATED ONLINE STUDIES

The choice measures described in Section “Examples and Replications I: Choice and Verbal Measures in Moderated Online Studies” are relatively straightforward to adapt online, but they can only be used with children who are old enough to follow verbal instructions. To explore how infants can be studied using moderated online methods, here we discuss ways to implement looking time measures, including preferential looking and violation of expectation (VoE) paradigms (see also Smith-Flores et al., 2021). As in Section “Examples and Replications I: Choice and Verbal Measures in Moderated Online Studies,” we review two sets of studies conducted in person and online implementing these measures. They demonstrate both the feasibility of conducting infant research online and that data collected online can yield comparable results to data collected in person.

Preferential looking is relatively straightforward to implement via moderated methods. It has traditionally been used to measure the preferences of infants who are too young to reach (e.g., Kinzler et al., 2007; Hamlin et al., 2010; Powell and Spelke, 2018); indeed, prior work has shown that younger infants often look at the same characters that older infants ultimately reach for. Preferential looking paradigms can be implemented online by presenting stimuli side by side and assessing the direction and duration of participants’ gaze; Study 3 presents an example implementation (see **Figure 3**). When using preferential looking as a dependent measure in online studies, infants’ positioning with respect to the camera’s field of view is important; although preferential looking studies can also be implemented using unmoderated methods, in moderated sessions, an experimenter can provide clear, real-time instructions to the parent about how best to position or reposition their child.

Violation of expectation paradigms (Aslin, 2007) can also be implemented online, using either unmoderated or moderated methods. Below we provide an example of a moderated online study (Study 4) where infants were shown sets of video stimuli for familiarization and then a test stimulus presented as a separate video; infants’ duration of looks at the test video was measured in the same way as an in-person VoE paradigm. Note that in online VoE studies, variability in camera angle, screen size, and video feed quality can make it hard for experimenters to determine when infants divert their gaze from the screen and for how long. Therefore, successful implementation of VOE studies in moderated sessions require a reliable method for tracking



FIGURE 3 | Screenshots of video stimuli implemented in Study 3 (preferential looking, see Woo and Spelke, 2020). **(A)** Participants were first familiarized to the bear's preferred toy. **(B)** The contents of the boxes were then switched either in the rabbits' presence or absence. **(C)** One rabbit opened the box where the desired toy was moved to while the other opened the one where the desired toy was originally. **(D)** At test, infants were shown the two rabbits. In person, infants were asked to choose the one they like and their reach was recorded; online, infants were presented with a video of the two rabbits and their preferential looking was measured.

infants' gaze and duration in real time. To address this issue in Study 4, two coders individually tracked and measured infants' looking duration². In general, variability across participants in experimental setup is an important factor to consider when deciding exclusion criteria tailored for online data collection.

In what follows, we present the key methods and results from two sets of studies with infants conducted in person and online. One examines 15-month-old infants' preferential reaching and looking (Study 3) and another examines 6- to 13-month-old infants' looking time (VoE, Study 4).

Study 3

Research Question

Do infants' social evaluations take into account the intentions of agents acting on false beliefs? The methods, analyses, results, and discussion of Study 3 were first reported fully in Woo and Spelke (2020). Infants viewed scenarios with social agents who possessed true or false beliefs about the outcomes of their actions directed toward another agent in need of help. Infants' preferential reaching (in-person) or preferential looking (online) toward the agents were measured. The primary research question in Woo and Spelke (2020) is whether infants prefer agents with helpful intentions or agents who cause positive outcomes.

Participants

In-person

Forty-six infants (23 females, mean: 15.06 months, range = 14;10–15;18) participated in person. An additional 15 participants

²Another solution is to use an infant-controlled looking procedure in which trial length is determined by an individual infant's looking behavior, but such procedures can be difficult to implement. For unmoderated studies, the Lookit platform has recently begun to support this feature.

were excluded, based on preregistered exclusion criteria (see Woo and Spelke, 2020, for full details about demographics and about exclusion).

Online

Forty-eight infants (26 females, mean: 14.91 months, range = 14;10–15;20) participated online, using Zoom's screen share features. Participants were recruited via local and online advertising. An additional four participants were excluded, based on preregistered exclusion criteria (see Woo and Spelke, 2020).

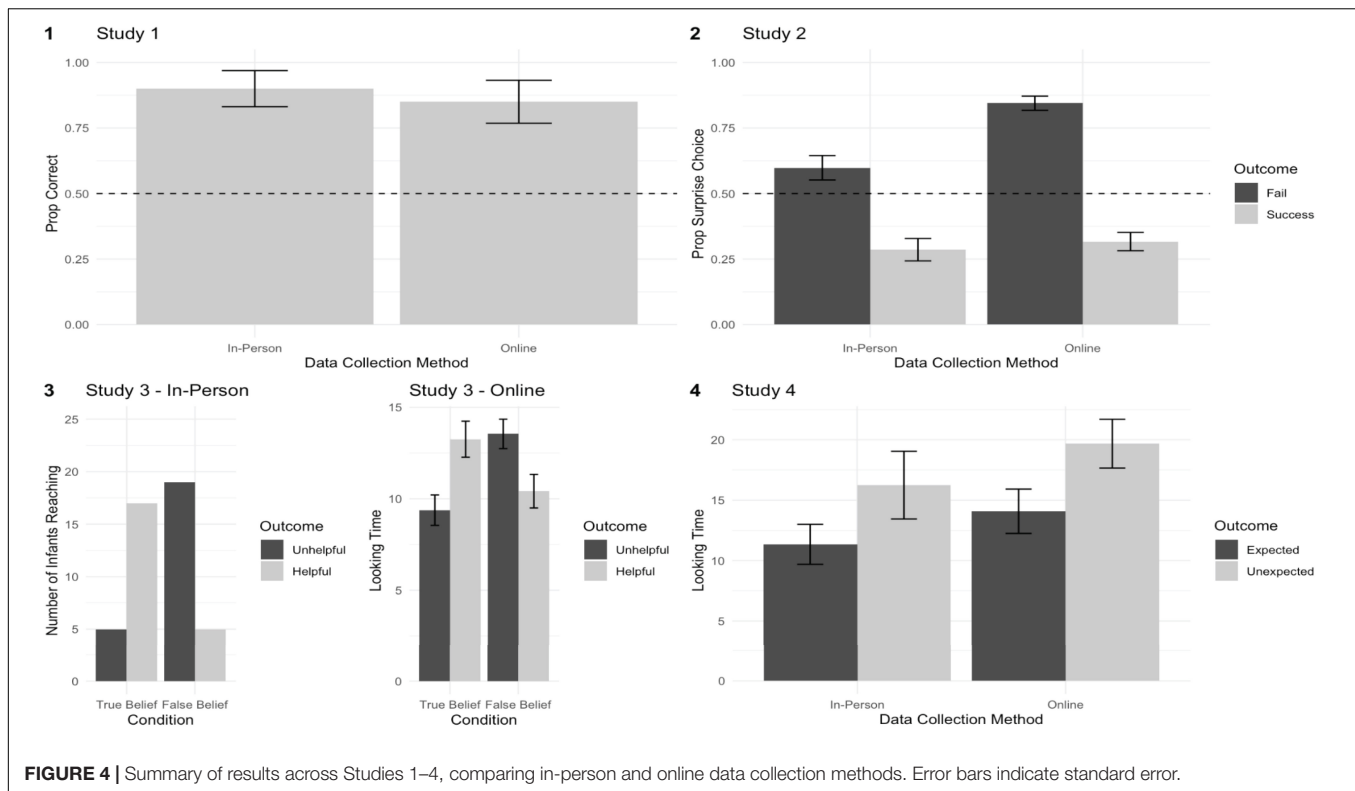
Method

Both in-person and online

Woo and Spelke (2020) familiarized infants to videos of puppet shows in which a bear protagonist repeatedly grasped a toy (its desired toy) in a box while two rabbits observed. Following familiarization, the toy was moved to a new box and both boxes were closed, either as the rabbits were present or absent. In their presence, the rabbits would have true beliefs about the location of the desired toy; in their absence, they would have false beliefs. In the final event, the bear returned. One rabbit opened the original box that had contained the bear's desired toy, whereas the other rabbit opened the new box that contained the bear's desired toy.

In-person

Infants sat on their caregiver's lap in the lab before a 102-cm by 132-cm LCD projector screen. After viewing all events, an experimenter assessed infants' evaluations through their preferential reaching behavior directed at the two rabbits. The experimenter determined the infant's choice as the first rabbit they touched via a visually guided reach (see Woo and Spelke, 2020, for reliability analyses).



Online

The online version of this experiment was almost exactly the same as in-person, except infants' evaluations were assessed through their preferential looking toward the two rabbits, presented via Zoom screen share. After the final event, the two rabbits appeared on opposite sides of the screen and moved to an experimenter's pre-recorded voice saying "Hi! Look! Who do you like?" three times, once every 10 s (see **Figure 3D**); infants' looking toward each rabbit was then assessed (see Woo and Spelke, 2020, for reliability analyses).

Results

In-person

Woo and Spelke (2020) found that, when rabbits had true beliefs about the desired toy's location, infants preferentially reached for the rabbit who opened the new box with that toy (17/22 infants, binomial $p = 0.016$, relative risk = 1.54). By contrast, when rabbits falsely believed that the desired toy was in its original box, infants reached for the rabbit who opened the original box (19/24 infants, binomial $p = 0.006$, relative risk = 1.58). The patterns of reaching based on outcomes (i.e., which box rabbits opened) differed significantly between conditions [$\chi^2(1) = 12.47$, $p < 0.001$, Wald's odds ratio = 12.92].

Online

When rabbits had true beliefs about the desired toy's location, infants preferentially looked at the rabbit who opened the new box with that toy [mean preference% = 58.2%, 95% CI [51.9%, 64.5%], SD = 14.9%, one-sample $t(23) = 2.71$, $p = 0.012$, $d = 0.55$]. When rabbits falsely believed that the desired toy was in its original box, infants instead preferentially looked at the rabbit

who opened the original box [mean preference% = 57.0%, 95% CI [50.8%, 63.2%], SD = 14.6%, one-sample $t(23) = 2.36$, $p = 0.026$, $d = 0.48$] (**Figure 3B**). Looking preferences based on outcomes differed significantly between conditions [two-sample $t(45) = 3.59$, $p < 0.001$, $d = 1.03$]. See **Figure 4 (3)** for a summary of results.

Study 4

Research Question

Do infants expect other agents to minimize the cost of their actions? The in-person version of this study was previously published as Experiment 1 in Liu and Spelke (2017). Infants were shown efficient and inefficient actions after a habituation (in-person) or familiarization (online) period, and their duration of looking toward those actions was measured. The hypothesis was that if infants expect an agent to perform an efficient action, then they will look longer when they perform an alternative, inefficient action³.

Participants

In-person

Twenty 6-month-old infants (10 females, mean age = 5.95 months, range = 5.6–6.3) participated in-lab. Seven additional infants were tested but excluded from the final sample (two did fussiness, one did not habituate, two because of experimental or technical error, and one for interference

³Further experiments helped constrain the interpretation that infants' looking preferences were not driven by interest in higher or faster jumps, and did not result merely from learning the relation between the height of the barrier and the height of the jump during habituation.

from caregivers) based on exclusion criteria specified ahead of data collection.

Online

The online replication included 27 infants ranging from 6 to 13 months of age (15 females, mean age = 10.4 months, range = 6.9–13.1). Participants were recruited via local and online advertising. Two additional infants were tested but excluded from the sample (one due to fussiness and one for failing to complete the test trial) based on exclusion criteria specified ahead of data collection. The age range for this sample was chosen to match or exceed the age of infants in the original study. The sample size was chosen based on a simulation-based power analysis, implemented using the *simr* package in R (Green and MacLeod, 2016), over the confirmatory analysis from the original study (comparison of looking time between the inefficient and efficient events).

Methods

Both in-person and online

Infants were first calibrated to the display screen using a toy that was held at the center, top, bottom, left, and right of the screen (in-person), or a video of an object that appeared at each of those locations (online). Parents were instructed not to engage with their infants or attract their attention toward or away from the stimuli. During the study, infants first saw looped videos of an agent leaping over a tall barrier. The height of this barrier varied slightly across loops, and the agent always conformed the height of its jump to the height of the barrier, following previous studies (Gergely et al., 1995). Then, at test, the barrier was removed and replaced with a lower barrier that infants had not seen before. Infants, on alternating trials, then saw the agent jump the same height as before (now inefficient) or jump low enough to just clear the new barrier (now efficient). The order of test events was counterbalanced across participants. All trials lasted until babies looked for 60s, or until they looked away for two consecutive seconds. The final data generated for the analysis was coded from video recordings by researchers who were naive about the order of test events presented to infants.

In-person

Infants sat on their caregivers' laps in front of a large projector screen. Infants saw 6 to 14 habituation trials, and 3 pairs of test trials. Infants met habituation criteria when their summed attention across the most recent 3 trials fell to below half of their summed attention across the first 3 trials, or after 14 habituation trials. For more details, see Liu and Spelke (2017).

Online

Infants viewed stimuli presented as a YouTube playlist on parents' laptop or tablet screens in their homes. They either sat on their caregivers' laps or in a high chair. An experimenter used Zoom's screen share and remote control features to move through the playlist and record the session. To simplify the study design, infants saw only six familiarization events and two pairs of test trials. The experimenter determined trial duration using *jHab* (Casstevens, 2007) and went to the next video in the playlist when indicated. Because of the variable screen sizes and setups across infants, and the lower quality of the videos from the study sessions, two naive coders generated the final

data rather than one; the initial coding took place during the session, and coders also reviewed the recordings afterward for accuracy. Neither had access to the view of the stimuli. If they disagreed about the duration of looking by more than 4s, or about whether a particular trial should be included or excluded, a third coder resolved the disagreement. In two cases, no video record of the session could be recovered, so the original coding generated during the experiment was used.

Results

Across both studies, the primary dependent measure was the average looking time toward the unexpected (i.e., inefficient action) vs. expected (i.e., efficient action) test event, log transformed to correct for skew in the data.

In-person

Infants looked longer at the familiar but inefficient jump versus the novel but efficient jump ($M_{\text{inefficient}} = 16.25$, $M_{\text{efficient}} = 11.35$, $[-0.49, -0.11]$, $\beta = -0.46$, $B = -0.3$, $SE = 0.1$, $p = 0.006$, two-tailed, mixed effects model with looking time in log seconds as the response variable, trial type as a fixed effect, and a random intercept for each participant, all subjects included based on a 4/n cutoff from Cook's Distance, where n is the number of participants (Nieuwenhuis et al., 2012).

Online

Infants who participated online performed similarly to those that participated in-person: infants looked longer at the inefficient jump versus the efficient jump ($M_{\text{inefficient}} = 19.68$, $M_{\text{efficient}} = 14.08$, $[-0.7, -0.14]$, $\beta = -0.611$, $B = -0.36$, $SE = 0.13$, $p = 0.012$, two-tailed, mixed effects model with looking time in log seconds as the response variable, trial type as a fixed effect, and a random intercept for each participant, removing one influential participant identified using Cook's Distance). Including this participant did not change the interpretation of the predicted effect. See **Figure 4 (4)** for a summary of results.

EVALUATING THE VALIDITY OF ONLINE DEVELOPMENTAL METHODS: A META-ANALYSIS

While the disruption of in-person testing was an unavoidable consequence of the COVID-19 pandemic, it also provided a rare opportunity to directly compare results across two nearly identical versions of studies administered online and in person. Collectively, the four studies presented in Sections "Examples and Replications I: Choice and Verbal Measures in Moderated Online Studies" and "Examples and Replications II: Looking Time Measures in Moderated Online Studies" successfully replicated the initial findings from in-person procedures using online (moderated) adaptations of the same procedures. In order to facilitate comparison of effect sizes across different studies using different dependent measures, we conducted a meta-analysis based on data from Studies 1–4. Meta-analysis is a standard method for aggregating effect sizes across disparate experimental paradigms (Hedges, 1992), and meta-analyses can even be effective ways to aggregate across small numbers of studies (Goh et al., 2016).

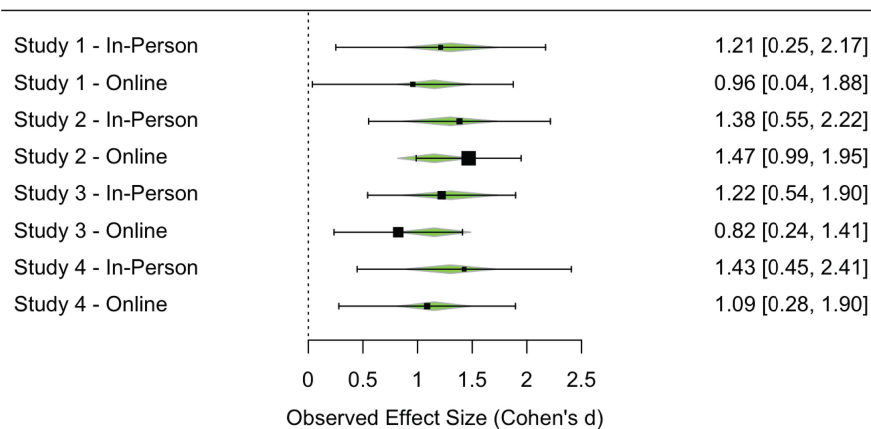


FIGURE 5 | Forest plot showing the standardized effect of interest across Studies 1–4, comparing in-person and online data collection methods. Points are sized based on sample, and error bars indicate effect size variance. Green triangles show random-effects multilevel meta-regression model estimates of the effect size for each study.

For each study reported here, we calculated the effect size and associated variance for the primary effect of interest, Cohen's d for Studies 3 (online) and 4 (in-person and online) and log odds for the remaining studies. We then converted all effects to Cohen's d and computed variance using the *compute.es* package in R (Del Re and Del Re, 2012). We then performed a random-effects multilevel meta-regression over the eight effect sizes using *metafor* (Viechtbauer, 2010). This meta-regression attempted to estimate the effect of online data collection, predicting this effect across the four pairs of experiments. A forest plot is shown in **Figure 5**. In aggregate, the meta-analysis estimated a small negative, non-significant effect of data collection method (online), $p = 0.58$, 95% CI $[-0.67, 0.38]$. This finding confirms the impression that, despite the differences in context and implementation, data collected in person and online elicited similar effects in the current studies, providing empirical support for the validity of moderated online data collection.

DISCUSSION

Online developmental studies are still in their infancy, and the idea of using video-conferencing tools for developmental studies may be new to many researchers. To help researchers decide how best to implement their own online studies, here we reviewed various considerations for software choice as well as techniques and strategies for designing effective studies that maximize participant attention and engagement. We then presented four examples of studies where an in-person experiment was replicated by adapting the procedures and stimuli for moderated online data collection.

Comparison between in-person and online studies suggests that moderated online data collection provides a viable alternative to in-person data collection. In Study 1, preschoolers' choice of agent in person was nearly identical to those who participated online. Similarly, in Study 2, elementary schoolers performance on the test question was significantly above chance in both the in-person and online versions of the study. In Study

3, infants' pattern of preferential reaching measured in person closely paralleled the pattern of infants' preferential looking measured online. In Study 4, infants' looking times across two conditions were comparable between the in-person and online versions of the study. Further, a meta-analysis revealed similar effect sizes across in-person and online data collection for the studies in the current sample.

Limitations

Although the overall results of the current studies suggest similar experimental outcomes for developmental studies conducted in-person and online, there are several factors that limit the generalizability of these findings and our ability to draw sweeping conclusions about online research as a whole. First, the current studies focus primarily on social cognition and therefore feature animated agents that exhibit various behaviors. The presence of such agents may have made these studies particularly interesting and engaging to infants and children (see Kominsky et al., 2020). Whether similar results would be expected in studies that only involve inanimate objects, shapes, or sounds remains an open question.

Second, the current studies utilize a small subset of possible measures (i.e., verbal choice, preferential looking, and looking time). The efficacy of other, more continuous measures, such as rating scales or free form responses, is less clear. Therefore, future research is needed to examine the viability and efficacy of a broader range of methods, measures, and research questions. Nonetheless, the current data suggest that the results of online developmental studies, when adapted properly, are comparable to those of similar studies conducted in person.

Third, these studies were conducted in the United States with participants who have relatively reliable internet access and are reasonably comfortable operating laptops, tablets, or smartphones. Therefore, the current results do not speak to the efficacy of online research in populations with unreliable internet access or less experience with telecommunications technology. Nonetheless, online data collection, in principle, offers easier access to some samples—particularly those in developing

countries with increasing internet-access—than traditional in-person data collection. Thus, we see online methods as a promising approach for improving the diversity, generalizability, and outreach of developmental research, and more broadly as an exciting direction of future research efforts that are larger in scale and impact (Sheskin et al., 2020).

Future Developments

Obviously, certain kinds of studies simply cannot be adapted to an online format, especially if they require special equipment or interaction with physical stimuli (e.g., neural measures, physical exploration tasks, etc.). Looking back at the past several years, however, many new strategies have been developed to collect various dependent measures using online data collection that were believed to be infeasible. We hope this trend continues, and look forward to new and exciting methods, measures, and research questions that can be implemented online. As online methods become more easily accessible and widely adopted, researchers across a greater variety of subdisciplines will adapt their studies to an online format. In turn, this will bring a greater variety of dependent measures and experimental paradigms. While our work provides preliminary support for some existing approaches, further research is needed to determine their efficacy compared to alternative approaches.

Online research allows researchers to easily expand the size and the demographics of their samples, with the potential to reach families across the world (Sheskin et al., 2020). In addition to the positive impact on the representation and generalizability of developmental research, this provides an unprecedented opportunity to improve community outreach and engagement. Although this can happen passively as more families participate in the scientific process, researchers can also actively update families on research findings and speak directly to those interested in developmental research. Conducting research online also makes it easier for students to get involved in the research process, especially those who may have limited access to universities with traditional in-person developmental research facilities. As resources for online research become more centralized, we hope more individuals take part in the scientific process as both researchers and participants.

CONCLUSION

Online developmental studies proliferated in part because of the COVID-19 pandemic, but they are likely here to stay. Here, we have described a number of countermeasures to the limits of the online medium, including ways to minimize the impact of technical issues and adapt developmental stimuli for online use. We also presented a meta-analysis of developmental studies conducted in-person and online, and found comparable results between both versions. Our main goal was to demonstrate the feasibility and promise of conducting developmental research online. With these initial steps, we hope that researchers continue to utilize the medium to innovate and improve the accessibility, diversity, and replicability of developmental science.

DATA AVAILABILITY STATEMENT

The data for Studies 1, 2, and 4 can be found in the following repository: <https://osf.io/rpvxw/>. See Woo and Spelke (2020) for Study 3 data.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Stanford University IRB, Harvard University IRB, MIT IRB. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

AC and HG drafted the initial manuscript. SB, BC, GD, TG, MM, SR, JS, NV, and XJZ helped develop online testing materials. MA, BC, JAL, SL, BW, and YW contributed the empirical data and drafted the study methods. AC and MCF conducted the meta analyses. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Jacobs Foundation, NSF (BCS-2019567), and James S. McDonnell foundation (HG) and the National Institute of Health (F32HD103363) (SL).

ACKNOWLEDGMENTS

We thank Elizabeth Spelke, Ashley Thomas, and the other members of the Harvard Lab for Developmental Studies for their contributions to developing methods for testing infants and toddlers online (providing the data for Studies 3 and 4). We also thank the Stanford Social Learning Lab for their time and dedication spent developing and testing methods for testing children online. We are grateful for the children and families who participated in our online studies as well as Bing Nursery School, Palo Alto Museum and Zoo, and Boston Children's museum for their support in collecting in-person data for Studies 1 and 2.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.734398/full#supplementary-material>

Online testing materials developed by the authors can be found at: https://github.com/sociallearninglab/online_testing_materials.

REFERENCES

- Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., and Rökkum, J. N. (2019). The MTurkification of social and personality psychology. *Pers. Soc. Psychol. Bull.* 45, 842–850. doi: 10.1177/0146167218798821
- Asaba, M., Wu, Y., Carrillo, B., and Gweon, H. (2020). “You’re surprised at her success? Inferring competence from emotional responses to performance outcomes,” in *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, eds S. Denison, M. Mack, X. Yang, and B. Armstrong (Austin, TX: Cognitive Science Society), 2650–2656.
- Aslin, R. N. (2007). What’s in a look? *Dev. Sci.* 10, 48–53.
- Bailenson, J. N. (2021). Nonverbal overload: a theoretical argument for the causes of Zoom fatigue. *Technol. Mind Behav.* 2.
- Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition* 21, 37–46. doi: 10.1016/0010-0277(85)90022-8.
- Cassteven, R. (2007). *jHab: Java Habituation Software (Version 1.0. 2)*.
- Del Re, A. C., and Del Re, M. A. (2012). *Package ‘compute.es’*.
- Gergely, G., Nádasdy, Z., Csibra, G., and Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition* 56, 165–193. doi: 10.1016/0010-0277(95)00661-h
- Goh, J. X., Hall, J. A., and Rosenthal, R. (2016). Mini meta-analysis of your own studies: some arguments on why and a primer on how. *Soc. Pers. Psychol. Compass* 10, 535–549. doi: 10.1111/spc3.12267
- Green, P., and MacLeod, C. J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods Ecol. Evol.* 7, 493–498. doi: 10.1111/2041-210X.12504
- Gweon, H., Asaba, M., and Bennett-Pierre, G. (2017). “Reverse-engineering the process: adults’ and preschoolers’ ability to infer the difficulty of novel tasks,” in *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, eds G. Gunzelmann, A. Howes, T. Tenbrink, and E. Davelaar (Austin, TX: Cognitive Science Society), 458–463.
- Hamlin, K. J., Wynn, K., and Bloom, P. (2010). Three-month-olds show a negativity bias in their social evaluations. *Dev. Sci.* 13, 923–929. doi: 10.1111/j.1467-7687.2010.00951.x
- Hedges, L. V. (1992). Meta-analysis. *J. Educ. Stat.* 17, 279–296.
- Kinzler, K. D., Dupoux, E., and Spelke, E. S. (2007). The native language of social cognition. *Proc. Natl. Acad. Sci. U.S.A.* 104, 12577–12580.
- Kominsky, J. F., Lucca, K., Thomas, A. J., Frank, M. C., and Hamlin, K. (2020). Simplicity and validity in infant research. *PsyArXiv [Preprint]* doi: 10.31234/osf.io/6j9p3
- Liu, S., and Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition* 160, 35–42.
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Nieuwenhuis, R., Te Grotenhuis, H. F., and Pelzer, B. J. (2012). Influence. ME: tools for detecting influential data in mixed effects models. *R J.* 4, 38–47. doi: 10.32614/rj-2012-011
- Powell, L. J., and Spelke, E. S. (2018). Human infants’ understanding of social imitation: inferences of affiliation from third party observations. *Cognition* 170, 31–48. doi: 10.1016/j.cognition.2017.09.007
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/opmi_a_00002
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (part 2): assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/opmi_a_00001
- Sheskin, M., and Keil, F. (2018). TheChildLab. com a video chat platform for developmental research. *PsyArXiv [Preprint]* doi: 10.31234/osf.io/rn7w5
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Smith-Flores, A. S., Perez, J., Zhang, M. H., and Feigenson, L. (2021). Online measures of looking and learning in infancy. *PsyArXiv [Preprint]* doi: 10.31234/osf.io/tdbnh
- Spelke, E. S., Breinlinger, K., Macomber, J., and Jacobson, K. (1992). Origins of knowledge. *Psychol. Rev.* 99:605.
- Stahl, A. E., and Feigenson, L. (2015). Observing the unexpected enhances infants’ learning and exploration. *Science* 348, 91–94. doi: 10.1126/science.aaa3799
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36, 1–48. doi: 10.3320/978-1-931504-81-2.1
- Woo, B. M., and Spelke, E. S. (2020). Infants’ social evaluations depend on the intentions of agents who act on false beliefs. *PsyArXiv [Preprint]* doi: 10.31234/osf.io/eczgp
- Wu, Y., Muentener, P., and Schulz, L. E. (2017). One-to four-year-olds connect diverse positive emotional vocalizations to their probable causes. *Proc. Natl. Acad. Sci. U.S.A.* 114, 11896–11901. doi: 10.1073/pnas.1707715114

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Chuey, Asaba, Bridgers, Carrillo, Dietz, Garcia, Leonard, Liu, Merrick, Radwan, Stegall, Velez, Woo, Wu, Zhou, Frank and Gweon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A New Look at Infant Problem-Solving: Using DeepLabCut to Investigate Exploratory Problem-Solving Approaches

Hannah Solby^{*†}, Mia Radovanovic[†] and Jessica A. Sommerville

Department of Psychology, University of Toronto, Toronto, ON, Canada

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Ori Ossmy,
New York University, United States
Bennett I. Berthenthal,
Indiana University Bloomington,
United States

*Correspondence:

Hannah Solby
hannah.solby@mail.utoronto.ca

[†]These authors share first authorship

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 04 May 2021

Accepted: 18 October 2021

Published: 08 November 2021

Citation:

Solby H, Radovanovic M and
Sommerville JA (2021) A New Look
at Infant Problem-Solving: Using
DeepLabCut to Investigate
Exploratory Problem-Solving
Approaches.
Front. Psychol. 12:705108.
doi: 10.3389/fpsyg.2021.705108

When confronted with novel problems, problem-solvers must decide whether to copy a modeled solution or to explore their own unique solutions. While past work has established that infants can learn to solve problems both through their own exploration and through imitation, little work has explored the factors that influence which of these approaches infants select to solve a given problem. Moreover, past work has treated imitation and exploration as qualitatively distinct, although these two possibilities may exist along a continuum. Here, we apply a program novel to developmental psychology (DeepLabCut) to archival data (Lucca et al., 2020) to investigate the influence of the effort and success of an adult's modeled solution, and infants' firsthand experience with failure, on infants' imitative versus exploratory problem-solving approaches. Our results reveal that tendencies toward exploration are relatively immune to the information from the adult model, but that exploration generally increased in response to firsthand experience with failure. In addition, we found that increases in maximum force and decreases in trying time were associated with greater exploration, and that exploration subsequently predicted problem-solving success on a new iteration of the task. Thus, our results demonstrate that infants increase exploration in response to failure and that exploration may operate in a larger motivational framework with force, trying time, and expectations of task success.

Keywords: cognitive development, exploration, infant development, motion capture technology, automated behavioral analysis, problem solving, DeepLabCut

INTRODUCTION

The ability to overcome obstacles to achieve one's goals is crucial to success across a broad range of contexts. Problem-solving is particularly ubiquitous early in life. Infants are faced with a multitude of new problems every day such as obtaining desirable, out-of-reach objects, navigating around barriers, and learning to operate new toys. Research suggests that infants typically adopt one of two approaches to solving problems: infants imitate the problem-solving solutions of others (e.g., Provasi et al., 2001; Esseily et al., 2010) or explore to generate their own solutions (e.g., Willatts, 1999; Fagard et al., 2014). However, the circumstances that influence whether infants adopt problem-solving approaches modeled for them versus explore new approaches are not well understood. In this paper, we investigate whether the nature of the social input infants receive

(namely, the effort and success of an adult's modeled problem-solving solution) and infants' own firsthand experience influence the degree to which infants use imitative versus exploratory problem-solving solutions.

Infants are capable learners and independently generate new solutions to problems via exploration. A variety of work suggests that when infants are presented with problem-solving paradigms in which they cannot obtain goal objects directly, they implement novel solutions (Goldfield, 1983; Willatts and Rosie, 1988; Babik et al., 2018). For instance, by the end of the first year of life, infants discover that they must reach or crawl around a barrier to retrieve a toy (Lockman, 1984; Lockman and Adams, 2001), and they can pull a cloth supporting an out-of-reach toy to get the toy (Sommerville and Woodward, 2005a,b). Critically, infants not only implement known solutions, but often explore new solutions iteratively until success is achieved (Willatts, 1990). With age, infants' ability to explore and innovate novel problem-solving solutions continues to improve. By 16 months of age, infants discover that they can use a rake as a tool to bring an out-of-reach toy into reach (Fagard et al., 2014). Thus, from an early age infants engage in exploration when presented with novel problems, often leading to problem-solving success.

Simultaneously, a variety of evidence suggests that infants also rely on imitation to solve problems. By 12 months of age, infants can already solve simple problems by replicating modeled solutions across a variety of contexts (Provasi et al., 2001; Brugger et al., 2007; Fagard and Lockman, 2009; Fagard et al., 2016). For example, 1-year-olds will learn from an adult model to grasp one end of a box while simultaneously raising a lid to overcome suction, to orient a bottle upside-down to retrieve a wooden peg, and to use a stick as a tool to retrieve a toy from a box (Esseily et al., 2010). Thus, infants readily imitate modeled solutions to facilitate their own success on novel problems.

While there is ample evidence that infants imitate, they do not do so indiscriminately; rather, infants remove superfluous components of modeled problem-solving solutions and explore their own solutions when modeled solutions are inefficient. For example, infants only replicate the exact actions of an adult model when they are the most efficient means to achieving a goal (e.g., Schwier et al., 2006). When steps are not causally necessary, infants are likely to skip these steps (e.g., Hauf et al., 2004; Brugger et al., 2007; Schulz et al., 2008). Not only do infants deviate from imitation by omitting superfluous steps, but infants also explore alternate solutions. When infants were shown a demonstration in which an experimenter turned on a light with their head with unconstrained hands, infants often utilized their own hands to turn on the light, achieving the goal more directly (Gergely et al., 2002; Zmyj et al., 2009). These studies demonstrate that there is some degree of fluidity in terms of whether infants will imitate versus explore when solving a problem.

In studies to date, imitation is often considered as qualitatively distinct from exploration, and consequently, researchers sometimes focus selectively on one approach or the other. For instance, Brugger et al. (2007) studied the effects of step necessity and adult modeling on imitation by coding whether infants performed two modeled steps. In this way, the authors successfully studied imitation, but exploration of novel solutions

was not considered. Likewise, many studies separately measure imitation and exploration by providing distinct definitions of each. For example, Muentener et al. (2018) investigated both exploration and imitation but devised separate tasks and scales to quantify each approach independently. At a global level, it is reasonable to consider the constructs of imitation and exploration separately given that researchers are trying to capture qualitatively distinct strategies. However, individuals often also engage in more nuanced explorations that involve variations on modeled solutions. For example, imagine an observer watches a model insert a key into a lock and turn it twice counterclockwise. After watching, the observer puts their own key into the lock, but it does not open. The observer may persevere in trying to reproduce the exact solution of the model, or they may enact a qualitatively distinct solution such as knocking on the door. On the other hand, the observer may also engage in micro-exploration: jiggling the key, varying force, or varying the angle they use to release the lock. When this variability is taken into consideration, it becomes clear that a continuous scale can be construed between faithful imitation and micro-exploration.

While micro-exploration has not been studied at the level of particular problem-solving strategies, there is evidence of micro-exploration within infant object interactions. Nuanced refinements of existing strategies have been argued to play a particularly important role in infancy both for acquiring motor skills (Robin et al., 1996; Keen, 2011) and for generating more complicated problem-solving strategies (Lockman, 2000). Specifically, to utilize tools and interact with objects, infants must learn to make subtle variations in their approaches, for instance by adjusting their trajectory and velocity while using a hammer (Kahrs et al., 2013) or by altering their grip on a food-laden spoon (McCarty et al., 1999), thus engaging in micro-exploration. These findings indicate that infants engage in micro-exploration to successfully handle objects and that they may also apply this ability to solve challenging problems.

Thus, the ability to integrate imitation and exploration into a continuum is important to understand the full spectrum of strategies that infants employ to solve problems. However, the ability to achieve this objective may be hampered by the inherent difficulty of quantifying imitation and exploration within a single objective and continuous scale. Research in this area has largely been accomplished through behavioral coding schemes and human raters. To do this, coders make qualitative judgments about whether a subject has imitated or explored, as well as the kind of exploratory strategy generated. In this way, participant behavior is coded and coerced into a discrete category structure. In order to investigate imitation and micro-exploration along a continuum, one must move beyond behavioral coding.

Fortunately, motion capture technology presents an avenue to generate continuous, objective measures of infants' motor responses in order to quantify infants' problem-solving approaches along an imitation-exploration continuum. Indeed, there is a rich history of using motion capture technology in developmental research to assess early motor, perceptual, and cognitive development (Thelen et al., 1996; Adolph et al., 2000; Berger and Adolph, 2003; Claxton et al., 2003; McCarty and Keen, 2005; Gill et al., 2009; Gottwald and Gredebäck, 2015;

Jung et al., 2015; Fragaszy et al., 2016; Gottwald et al., 2017, 2019; Ingvarsdóttir and Balkenius, 2020). Advancements in artificial intelligence have expanded access to motion capture by creating free, online programs for *post hoc* analysis such as frame difference (e.g., Paxton and Dale, 2013) and computer vision methods (e.g., Ossmy et al., 2020; Cao et al., 2021). While these computer vision methods have existed for the last decade, they have not been broadly employed in the field of developmental psychology and are not yet in the toolbox of most developmental researchers.

Here, we focus particularly on DeepLabCut (DLC) which allows users to train a neural network to track motion in up to three dimensions. DLC relies on a specialized algorithm which is pre-trained (Deng et al., 2009) such that DLC's neural network only requires a small number of frames for training and can manage lower resolution footage (Mathis et al., 2018; Cronin et al., 2019). As such, DLC is suitable for use with small samples but provides quality comparable to even commercial systems (Sturman et al., 2020; Vonstad et al., 2020). Further, DLC is incredibly versatile and is even able to track multiple, distinct individuals (Mathis et al., 2018, 2020). Thus, once trained, a researcher can utilize DLC with diverse data sets and variables of interest. Finally, the software is open-source, and its use has rapidly expanded across disciplinary lines in the last 3 years (for a scoping meta-analysis, see **Table 1**). The open-source nature of the program has stimulated an online community, with researchers introducing specialized packages (e.g., Fiker et al., 2020; Forys et al., 2020). As a field, developmental psychology has a history of active contribution to open-source projects, with many researchers releasing specialized packages in programming languages for others' use (e.g., Burke, 2019; Kominsky, 2019; Sanchez et al., 2019). As such, while developmental psychologists seem largely unaware of this technology, they are well-positioned to benefit from the rich and accurate behavioral data generated by DLC, as well as to contribute to the larger DLC community.

The goal of the current paper was to apply DLC to archival videos (Lucca et al., 2020) to capture infants' problem-solving approaches to a challenging problem. In the original study, Lucca et al. (2020) provided 18-month-old infants with a modeled solution to a means-end problem: infants watched an adult experimenter pull a rope that was attached to an out-of-reach transparent box containing a toy, in order to bring the box and toy within reach. Infants saw one of three demonstrations that varied in terms of effort and success. In the Easy condition, the experimenter pulled the rope, and the box immediately came within reach, allowing her to retrieve the toy. In the Hard condition, the experimenter pulled the rope five times. On the first four pulls, the box did not move despite the experimenter's efforts. On the fifth pull, the box slowly began to move until it was completely brought into reach, allowing her to retrieve the toy. The Impossible condition demonstration was similar to the Hard condition, except that the experimenter never succeeded in moving the box and thus was unable to retrieve the toy. After observing the demonstration, infants were presented with an impossible test trial in which the toy was surreptitiously affixed to the table. This cycle was repeated three times, and the researchers measured how long infants engaged in pulling the rope to retrieve

the toy, as well as negative affect, maximum pulling force, help-seeking, and hints required during a subsequent recovery trial (designed to test supported needed on a new iteration of the task).

Lucca et al. (2020) found that the effort and success of the adult model and accumulating firsthand experience with failure jointly influenced how long infants attempted to solve the problem, as well as several measures of performance. For instance, trying time dramatically decreased across trials in the Easy condition as infants experienced greater firsthand failure. Here, the success of the experimenter model suggested that infants should succeed quickly by employing the experimenter's approach. As such, infants may have inferred that they did not have adequate skill to solve the problem. Similarly, in the Impossible condition, trying time also dramatically decreased across trials, but also started off relatively low. In this case, the experimenter's failure, coupled with firsthand failure across trials, may have led infants to infer that the task was simply impossible. On the other hand, in the Hard condition, infants' efforts remained relatively stable across trials. In this case, infants' inferences about the problem were presumably influenced by both sources of information. The experimenter demonstration suggested that the problem was solvable but difficult, requiring infants to try for sufficiently long to succeed. Thus, the firsthand failure infants experienced was not surprising or demotivating, leading to continued trying despite failure. Together these findings indicated that the effort and success of the adult model, along with accumulating firsthand experience with failure, influence how long infants try to solve a given problem.

Thus, the current study had three objectives. First, we investigated how two manipulated factors, the effort and success of an adult model's problem-solving solution and firsthand experience with problem-solving failure, influenced the degree to which infants adopt imitative versus exploratory approaches. Second, we looked at how individual differences in other performance measures on the task, such as infants' negative affect, maximum pulling force, help-seeking, and trying time predicted infants' exploration. Finally, we were interested in if imitative versus exploratory approaches predicted motivation on a functioning version of the task.

To address our three objectives, we trained DLC to track the coordinates of the rope handle the infants pulled during the problem-solving task. Specifically, we considered imitation in this context to have two components: (1) visible similarity to the approach employed by the experimenter, and (2) consistency of employment across time. Thus, we examined how model success and firsthand experience with failure influenced the degree of infants' *imitative similarity* to the experimenter (i.e., the extent to which infants copied the experimenter model), as well as the *variability* in their attempted solutions (i.e., how much infants varied the location of their attempts). Imitative similarity was measured using the displacement of rope pulling in the *x*- and *y*-axes relative to imitative pulling, and variability was measured using each participant's standard deviation of spatial displacement. Within the context of these variables, a decidedly imitative problem-solving approach would be marked by high imitative similarity and low variability, while a decidedly exploratory problem-solving approach would be marked by

TABLE 1 | A breakdown of peer-reviewed studies that have used DeepLabCut, organized by field of study, between 2018 and 2021.

| Field of study | | Number of articles | Authors (year) |
|---------------------------|--------------------------|--------------------|---|
| Agriculture | | 2 | Liu et al. (2020); Fang et al. (2021) |
| Biology | | 1 | Ho et al. (2021) |
| Biomechanics | | 2 | Cronin et al. (2019); Cronin (2021) |
| Neuroscience/Neurobiology | | 12 | Mathis et al. (2018); Wei and Kording (2018); Fiker et al. (2020); Forsys et al. (2020); Fried et al. (2020); Kim et al. (2020); Mathis and Mathis (2020); Mundorf et al. (2020); Rodriguez et al. (2020); Sturman et al. (2020); Whishaw et al. (2020); Williams et al. (2020) |
| Orthopedics | | 1 | White et al. (2020) |
| Physiology | | 4 | Barrett et al. (2020); Haberehner et al. (2020); Wu et al. (2020); Brandt et al. (2021) |
| Psychology | Comparative Psychology | 1 | López Pérez et al. (2021) |
| | Developmental Psychology | 0 | |
| | | | |
| Science and Technology | | 5 | Nath et al. (2019); Mathis et al. (2020); Vonstad et al. (2020); Huang et al. (2021); Namba et al. (2021) |

In order to identify the scope of DLC in different fields, we gathered publications that used DLC by using PsycINFO and the first 15 pages of Google Scholar with the search term "DeepLabCut." This analysis revealed 30 peer-reviewed papers that used DLC since its inception in 2018. Of these 30 papers, zero were in the field of developmental psychology. While DLC is a state-of-the-art software and is widely acknowledged in other fields such as neuroscience (Fried et al., 2020), developmental psychology has not yet taken advantage of this software.

low imitative similarity and high variability as multiple, novel solutions would be tested.

Regarding our first research question, it is plausible that we would find a pattern similar to Lucca et al. (2020), wherein infants would respond to the effort and success of the adult model, and their firsthand failure. A successful model (Easy condition) suggests that infants should imitate the solution for similar success, consistent with prior work showing that adults and children tend to favor imitating the solutions of others when others are successful (Schulz et al., 2008; Rendell et al., 2010; Wisdom and Goldstone, 2011; Reindl and Tennie, 2018). On the other hand, when the model fails (Impossible condition), infants may be more likely to explore because they think imitation is unlikely to solve the problem. Indeed, children and adults also increase rates of exploration when modeled examples are lower quality or less reliable (Rook and van Knippenberg, 2011; Carr et al., 2015). When the adult shows it is difficult but possible to solve the problem (Hard condition), infants' responses may fall between these two possibilities.

However, it may be the case that infants are influenced mostly by their firsthand experiences with failure, given that infants uniformly experience firsthand failure in all conditions and trials. Prior work has indicated that children increase exploration when the success of outcomes is unclear or surprising (Schulz and Bonawitz, 2007; Gweon and Schulz, 2008; Stahl and Feigenson, 2015; Bridgers et al., 2019). Thus, given infants' consistent experience with failure during the test trials, we expected that as a group, infants would explore solutions different from the experimenter, and that greater experience with failure (within trials and across trials) would decrease imitation, as continued failure would suggest imitation was not fruitful.

In order to address the second research question, we investigated whether negative affect, maximum pulling force, help-seeking, and trying time predicted infants' imitative

versus exploratory approaches. It is possible infants' affective responses may drive exploratory approaches, as infants may be more likely to abandon modeled solutions when frustrated. Similarly, infants may be more likely to adopt exploratory approaches when they have exerted maximal force when pulling the rope, compared to when they have only used minimal force. Addressing whether help-seeking predicted exploratory approaches will shed light on the extent to which the use of micro-exploration is predicted by the adoption of qualitatively distinct approaches. Furthermore, investigating whether trying time predicts exploratory approaches will inform whether these two metrics of performance signal conceptually related phenomena (i.e., different forms of persistence) or distinct phenomena.

Finally, we investigated whether exploratory approaches during test trials predicted hints needed during recovery trials when the task was solvable. While imitative and exploratory approaches are both means to remain engaged on the rope-pulling task, infants may generate different expectations through engagement in each approach. If a decidedly imitative approach is adopted, infants will uniformly experience failure every time they employ their method. This experience would likely lead to low expectations that the method will succeed the next time it is employed. On the other hand, each exploratory strategy infants test presents a possibility of success, even if small. Thus, infants who try a variety of strategies may require less support on a new iteration of the task.

MATERIALS AND METHODS

Participants

Participant videos from Lucca et al. (2020) were repurposed for this study. In the original study, 96 full-term, typically developing 18-month-olds (38 females, mean age = 18.50 months,

range = 17.67–19.30 months) participated. Participants had previously signed up to partake in studies through a university database and were recruited through this database for this study. Participants were parent-reported as White ($n = 69$), Asian ($n = 3$), Hispanic ($n = 2$), mixed race ($n = 21$), or declined to report ($n = 1$). The sample size was limited to that of the original study.

Motor Skills Checklist

In order to ensure our results were not constrained by individual differences in motor coordination, a measure of gross motor development was administered. Parents were given a 24-item motor ability checklist (Loucks and Sommerville, 2013) that was a variation of the Bayley Scales of Motor Development (Bayley, 2006). Questions pertained to infant's motor abilities and were organized in chronological order of developmental milestones (e.g., "Can your child sit alone while playing with a toy?," "Does your child attempt to walk?," "Can your child stand on one foot with help?"). The highest consecutive item parents checked served as a measure of motor development.

Procedure

Lucca et al.'s (2020) procedure consisted of three components: (1) a warm-up to familiarize the infants to their new environment, (2) demonstration-test trials: the experimenter first tried to retrieve an out-of-reach toy, then the infant was given the opportunity to retrieve the out-of-reach toy (this cycle was repeated three times with three different toys), and (3) a recovery trial. A more detailed description of the methods can be found in the original publication (Lucca et al., 2020); here we highlight the most important components for the current study.

During the demonstration-test trials, caregivers were seated and wearing occluding eyeglasses while infants sat on their caregiver's lap. In the demonstration phase, the infant observed the experimenter attempt to retrieve an out-of-reach toy in a transparent container by pulling on a rope that was attached to the container. Infants were assigned to one of three conditions. In the Easy condition, the experimenter easily retrieved the toy by pulling on the rope. In the Hard condition, the infant saw the experimenter struggle (she pulled the rope five times, and the box did not move), and eventually succeed at retrieving the toy by pulling on the rope. In the Impossible condition, the infant saw the experimenter try to pull the rope to retrieve the toy the same number of times as in the Hard condition, but she did not succeed.

During the test phase, infants were presented with the same toy in a container, attached to a rope. Unbeknownst to infants, the apparatus had been replaced with an identical looking version such that the container was stuck to the tabletop making the problem impossible to solve. Each trial ended after 120 s total had passed, or if the infant had not touched the rope for 15 s. The demonstration-test sequence was repeated three times, with three different toys. We analyzed the first 20 s of each trial for two reasons. First, as we wanted to adhere closely to the original methods. Second, as trials were variable in length and we believed it was important to have an equal amount of data from each participant for inferential purposes (i.e., data from 120 s

would likely be non-representative as it would only reflect the most active infants). As such, our sample would experience rapid attrition if other cut-offs were used (e.g., only 74% of participants had trials which each lasted at least 30 s; see **Table 2**).

Finally, infants participated in the recovery trial in order to observe infants' expectations of task success after the demonstration-test trials. Infants were again faced with a toy in a clear container attached to a rope; this time the apparatus was functional.

Coding

We focused on select variables coded by Lucca et al. (2020); namely, time spent trying and maximum pulling force. Additionally, data were collected on help-seeking behaviors, and affect. During the recovery trial, the number of hints were recorded.

Time Spent Trying

A primary coder watched each participant video and recorded the number of seconds the infant spent trying. An infant was classified as trying if they pulled the rope and looked directly at the toy immediately prior to, during, or after pulling. Behaviors such as swinging the rope side-to-side without making eye contact with the toy, or throwing the rope were classified as off-task behaviors, and therefore, not coded. A secondary coder independently double-coded 100% of the videos, establishing high reliability ($ICC = 0.95$, $p < 0.001$). The trying time data

TABLE 2 | Participant attrition in the Lucca et al. (2020) sample using different cut-offs for trial lengths.

| Trial time cut-off | % participants retained |
|--------------------|-------------------------|
| 20 s | 100% |
| 25 s | 90% |
| 30 s | 74% |
| 35 s | 55% |
| 40 s | 49% |
| 45 s | 41% |
| 50 s | 33% |
| 55 s | 32% |
| 60 s | 27% |
| 65 s | 25% |
| 70 s | 25% |
| 75 s | 23% |
| 80 s | 17% |
| 85 s | 15% |
| 90 s | 13% |
| 95 s | 10% |
| 100 s | 9% |
| 105 s | 8% |
| 110 s | 7% |
| 115 s | 6% |
| 120 s | 4% |

The percentage of participants whose trials each lasted a minimum length rapidly declines after 20 s.

was not normally distributed, therefore the data were square root transformed for analyses of trying time.

Maximum Pulling Force

The strength of trying was quantified using a 5 kg S-type load cell discretely connected to the toy. The load cell measured each infant's pull in pounds per square inch (PSI) and recorded continuous force data on a connected laptop. Maximum PSI was extracted during the first 20 s of each test trial. The force data was not normally distributed, therefore the data were also square root transformed for analyses of force.

Help-Seeking Behaviors

A primary coder watched each participant video and tallied the number of help-seeking behaviors displayed during the test trials. Help-seeking behaviors were defined as (1) reaching to the target object, or (2) points toward the target object or experimenter. Since parents were instructed to wear occluding eyeglasses, behavior directed toward the caregiver objectively could not be informative in this task, thus these behaviors were not coded as help-seeking. A secondary coder independently double-coded 100% of the videos, establishing strong reliability ($ICC = 0.93$, $p < 0.001$). For analyses, a composite help-seeking measure was created by summing both reaching and pointing behavior which occurred during each trial.

Affect

Participant affect was coded during bouts of trying, using still frames sampled at every 15 frames (i.e., every 510 ms) during trying time. Coders watched close-up recording on each participant's face and coded emotional reactions using a coding scheme adapted from Repacholi et al. (2016). Coding of positive and negative affect was performed separately, therefore in very rare instances infants could be coded as displaying both negative and positive affect during a single frame. Positive affect was coded if infants displayed characteristic features of a smile (e.g., upturning of the mouth, cheek elevation, raised brows). A secondary coder independently double-coded 50% of the videos for positive affect, establishing high reliability ($ICC = 0.97$, $p < 0.001$). Negative affect was coded if infants displayed characteristic features of frustration or disgust (e.g., down turning of the mouth, furrowed brows, wrinkled nose). A secondary coder independently double-coded 50% of the videos for negative affect, establishing high reliability ($ICC = 0.96$, $p < 0.001$). The total number of frames during a trial that the infant displayed negative affect was used in analyses.

Number of Hints Needed During Recovery Trial

The number of hints the infant required from the experimenter to complete the task during the recovery trial were coded. Unlike test trials, task success was possible during recovery. Therefore, hints required was used as a proxy to assess infants' expectations of task success (i.e., more hints would relate to greater expectations of failure). The number of hints provided by the experimenter was strictly a function of the amount of time passed, as hints were provided at fixed intervals if infants did not solve the task independently. Thus, hints did not reflect the infants' behavior or help-seeking. A secondary coder

independently double-coded 50% of the videos for hints required during recovery, establishing high reliability ($ICC = 0.96$, $p < 0.001$).

DeepLabCut Coding and Processing

As with our other variables, participant videos were trimmed to the first 20 s of each trial to match the hand-coding scheme employed by Lucca et al. (2020) and to ensure trials had equal data. However, given the variability in trial lengths, we also controlled for the trial length in our models to account for differences in overall time trying between participants. To this end, a human coder classified the length of trials to the nearest second. A secondary coder independently double-coded 25% of the videos for trial length, establishing high reliability ($ICC = 0.98$, $p < 0.001$).

In order to quantify infants' motion through space, DeepLabCut (DLC), a markerless pose estimation software, was applied to generate data on the rope coordinates for each participant in the x - and y -axes. Data were not collected in the z -axis as this is where pulling by the experimenter model occurred; as such, motion in the z -axis was considered imitative, rather than exploratory. The coordinates in the x - and y -axes were measured in units of pixels and represented the displacement of the rope relative to the origin (i.e., the top left corner of the camera view). Coordinates were generated at a rate of one coordinate pair per video frame (i.e., 30 times per second).

Our goal was to train an artificial neural network (ANN) to identify and track the rope handle through space. The first step was to hand-label frames with our points of interest to create training and test data sets that would be used to train the ANN. Estimation accuracy of a network is improved when trained to track more than one point (Mathis et al., 2018). As such, we trained the network to track three points on the rope: the beginning of the rope handle, the middle of the rope handle, and the end of the rope handle (see Figure 1). However, analyses were conducted using the beginning of the handle, where it attached to the rope. After observing labeled participant videos, this point was the least likely of the three to be occluded by the infants' hands, and thus, was the most reliable. DLC boasts < 5 -pixel error when trained on 100 hand-labeled frames (Nath et al., 2019). In order to minimize the pixel error on our data set, we labeled 200 frames from 10 participant videos. We chose 10 participant videos that represented the diversity of the sample both in terms participant demographics (e.g., participant race, participant gender), and in terms of perceptual features (e.g., shirt color) to ensure the training could be applied to the versatile range of participants present in Lucca et al.'s (2020) sample. Then, 200 frames that featured infants in a variety of positions were manually selected from the 10 participant videos.

Using the recommended neural network, ResNet-50 (Nath et al., 2019), we trained the network on 200,000 iterations of the training set. The trained network had a mean training error of 2.01 pixels, and a mean test error of 3.07 pixels (less than a quarter of a cm). We found this mean error size suitable for our work, thus we did not generate additional iterations of training and the remainder of the participant videos were analyzed using this network. For a detailed user guide, including instructions on

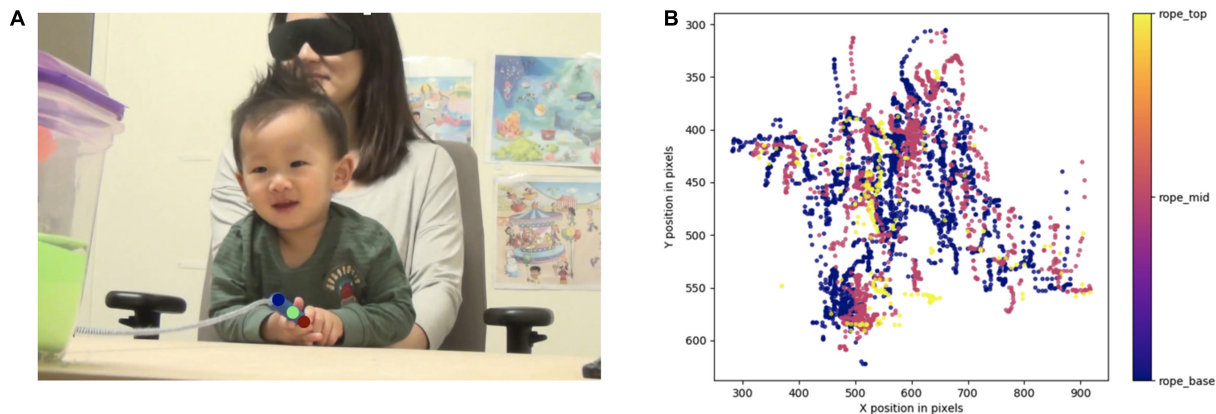


FIGURE 1 | (A) A still from a participant video showing markers generated by DLC after training the neural network. The three parts of the rope labeled and tracked were: the base of the rope handle (blue), the middle of the rope handle (green), and the end of the rope handle (red). **(B)** A graphical representation of the coordinates extracted from DLC. Additionally, we have uploaded a video of one of this participant's test trials with computed marker overlay. Video playback is in real time and can be found here: <https://osf.io/5z74k/>.

creating a training data set, training the network, and evaluating the trained network, please see Nath et al. (2019).

Assessing DeepLabCut's Precision

To assess the precision of DLC's labeling capacity, we had two human coders evaluate the labels generated by DLC. We randomly selected 25% of participants and then randomly selected 20 frames from each participant for evaluation (as 20 frames per participant were used in our training set). We found that on 90% of frames, DLC reported that it had detected the rope handle accurately; on 97% of those frames, human coders agreed that it was correctly labeled. For comparison, Sturman et al. (2020) found DLC was $86 \pm 3\%$ accurate compared to human annotated frames, and this outperformed commercial solutions.

Assessing DeepLabCut's Validity

Once data was extracted from DLC, it was utilized to construct several measures of exploration (see "Results" section). Before conducting analyses, we wanted to verify that our measures of exploration were not merely a reflection of low-level motor phenomena. To this end, behavioral coding was performed by human raters to identify times when infants were engaged in playing and unproductive movements (i.e., times when the rope was being moved but was not taut). Because infant's attempts were generally stochastic, shifting rapidly from one behavior to another, coding was performed on the level of 5-s intervals to allow for consistent classification between human raters. Two coders were assigned approximately half the sample each and marked how many intervals displayed unproductive movement in each trial for each participant. Thus, infants could score up to four intervals as unproductive per trial. In addition to coding assignments, approximately 25% of data was double coded and interrater reliability was moderately high ($ICC = 0.83, p < 0.001$).

To understand whether our measures of exploration inadvertently captured incidental movement, rather than concerted trying, we performed Pearson correlations between

our trial-level measures of imitative similarity and variability with the number of intervals engaged in unproductive movements. Indeed, we did not observe a correlation between either average imitative similarity ($r = -0.03, p = 0.57$) or overall variability ($r = -0.02, p = 0.68$) and our human-generated measure of unproductive movement. Thus, we did not find evidence that our measures of exploration reflected off-task behaviors or play. This verification provided us with increased confidence of the construct validity of our measures of similarity and variability as exploratory problem-solving strategies.

RESULTS

Effects of Firsthand Experience and Model Success on Infants' Imitative Similarity

In all the demonstrations, infants witnessed the experimenter modeling straight back pulling of the rope. Our first goal was to understand whether infants' trying attempts were similar to the demonstration of the experimenter or whether infants altered the angle of their pulling along the x - and y -axes. To this end, we produced measures to capture divergence from imitation by centering participants' raw displacement values in each axis relative to imitative (i.e., straight back) pulling. As not all participants engaged in imitative pulling, we were unable to center each participant's attempts relative to their own imitation. However, data from all participants who engaged in imitative pulling ($n = 78$) were utilized to generate average imitative estimates. A given pull was defined as imitative if the rope handle did not go beyond the shoulders in either axis, and if the infant was properly seated in their parents' lap (i.e., not straining or bouncing). These video clips were run through the neural network to obtain an average x -value (458.29 pixels; 35.25 cm from the left of the camera frame) and an average y -value (347.79 pixels; 26.75 cm from the top of the camera frame) for

imitative pulling. These values were subtracted from infants' raw displacement values for each axis. Because we valued divergence from imitation in both directions (i.e., right and left, up and down), the absolute value of each deviation value was then taken.

Once these values were calculated, we performed two one-sample *t*-tests to compare the displacement values in each axis to imitation (i.e., 0) to understand whether infants' pulling attempts differed significantly from the experimenter. To make use of the rich data produced by DLC, each *t*-test considered 60 points per participant (20 coordinate pairs per trial for each of the three trials), excluding outliers.¹ In each of the conditions, infants experienced failure once they attempted to solve the means-end problem on their own. Thus, we expected that infants as a group would generate new strategies to improve upon the strategy modeled by the experimenter. Pulling attempts in both the *x*-axis [$M = 191.01$ pixels/14.69 cm, $SE = 1.63$ pixels, $t(5670) = 117.09$, $p < 0.001$] and the *y*-axis [$M = 154.08$ pixels/11.85 cm, $SE = 1.16$ pixels, $t(5751) = 133.30$, $p < 0.001$] differed significantly from imitation. Thus, infants' attempts differed significantly from the experimenter in each axis. We also sought to understand whether pulling attempts differed between the two axes, thus we additionally performed a paired-sample *t*-test to understand whether there was greater deviation in one axis than the other. Indeed, infants' pulling attempts in the *x*-axis deviated from imitation to a significantly greater extent than in the *y*-axis [$M_{diff} = 37.36$ pixels/2.87 cm, $t(5669) = 18.67$, $p < 0.001$]. Thus, infants did not merely replicate the actions of the experimenter and their attempts appeared to differ to a greater extent in the *x*-axis than the *y*-axis.

Our next goal was to understand how infants' imitative similarity was influenced by the effort and success of the adult model and firsthand experience with failure. To this end, two measures were constructed using the absolute values from the *x*- and *y*-axes: (1) a difference score to allow us to understand if infants systematically varied the axis of their exploration, made by subtracting values in the *y*-axis from the *x*-axis, and (2) an additive score representing overall imitative similarity by summing, then reverse-scoring, the scores for interpretability. Thus, for the difference score, positive values indicate greater deviation from imitation in the *x*-axis than the *y*-axis, and for the imitative similarity measure, a score of 0 indicated imitative pulling in both axes and greater negative values represent greater exploration (see **Figure 2**). Once these measures were calculated, linear mixed effects models were built to predict changes in the two measures, respectively. In each model, participants were entered as random effects and the main effects of condition, trial number, and time within the trial (in seconds) were entered as predictors. Further, individual variation in overall trial length and motor skill may lead to differences in infants' experiences trying in this task. Therefore, we also entered the main effects of overall trial length and motor skill as covariates to control for these effects. As we expected that greater experience with failure (both within trials and across trials) would decrease the utility of imitation, we additionally checked for an interaction between

trial number and time within the trial. Finally, as condition was a categorical variable with three categories, we used the Easy condition as a baseline in accordance with Lucca et al. (2020), though it is worth noting that the pattern of results is the same regardless of specified baseline condition.

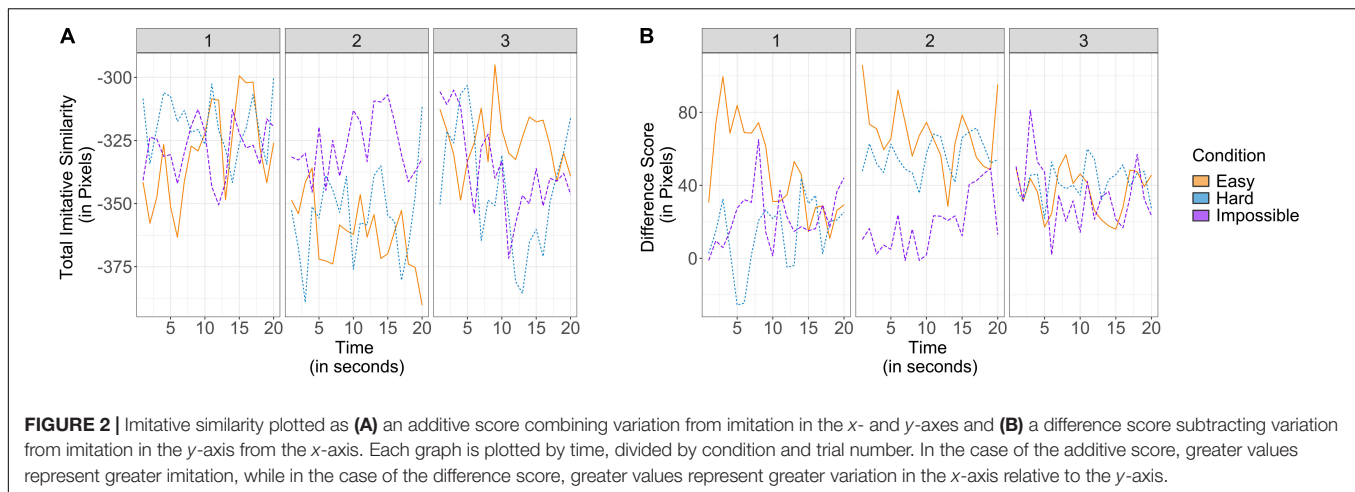
In the case of our difference score, we did not expect to find systematic variability as there was no information provided by the experimenter or across time which would suggest exploration in one axis would be more effective than in the other. Indeed, there were no significant main effects of condition, time within trial, nor an interaction between trial number and time (all p 's > 0.39). However, there was an effect of trial such that infants' pulling attempts differed from imitation to a greater extent in the *x*-axis than the *y*-axis in later trials [$t(5602) = 2.27$, $p = 0.02$, $\beta = 8.89$, $SE = 3.92$]. Thus, infants appeared to explore locations that were more disparate from the experimenter particularly in the *x*-axis across trials. This result may be due to the physical limitations of the study design, wherein infants sat in their caregiver's lap and were less able to move vertically than horizontally. Critically, there was not a significant effect of motor skill on the difference score ($p = 0.96$).

On the other hand, we thought that our measure of imitative similarity could be sensitive to the effort and success of the adult model, as infants received varying information about the success of the modeled solution, and to firsthand evidence, as failure would suggest a necessity for strategy diversification. Our model of imitative similarity revealed a significant main effect of trial such that infants' attempts became more similar to the experimenter over trials [$t(5497) = 2.14$, $p = 0.03$, $\beta = 8.09$, $SE = 3.78$] and a main effect of time such that infants' attempts became more imitative as trials progressed [$t(5475) = 2.51$, $p = 0.01$, $\beta = 1.63$, $SE = 0.65$], as well as a significant interaction between trial number and time such that on later trials, infants pulling attempts diverged more from imitation over time [$t(5475) = -3.08$, $p = 0.002$, $\beta = -0.93$, $SE = 0.30$]. As before, we did not observe an effect of motor skill in our model of imitative similarity ($p = 0.22$). It is worth noting that the effects in this model were relatively small, and that our prior analyses revealed infants' pulling attempts were overall significantly different from the experimenter in both axes. Thus, though infants' pulling attempts became more imitative over time, these attempts were still overall dissimilar to the experimenter. Finally, we did not observe any effects of condition on imitative similarity (both p 's > 0.78). Thus, imitative similarity seemed to respond more to information gained through firsthand experience than from the adult model.

Effects of Firsthand Experience and Model Effort and Success on Infants' Variability

Our analyses of imitative similarity allowed us to understand whether the locations of infants' pulling attempts varied significantly from the experimenter and how they varied over time. However, in the face of continued failure, it is both sensible to divest from imitation and also to test multiple locations and solutions as each new attempt fails. Thus, our next goal was to

¹In all analyses, outliers which were more than 2.5 SD from the mean were removed using pairwise deletion.



complement our understanding of exploration by evaluating the variability in infants' pulling attempts. To index spatial variability, the coordinates returned from DLC for each participant and each trial were used to calculate standard deviations of displacement in both axes. Because standard deviation is highly dependent on the mean of a given time interval, we calculated two variability scores: (1) a per-second variability score, representing the average standard deviation of movement in the x- and y-axes during the previous second, which responded to the local means of displacement in the previous second, and (2) an overall variability score, representing the average standard deviations of displacement in the x- and y-axes during the trial, which responded to the global means of displacement over the entire trial (see **Figure 3**). Because of skew in the per-second variability score and for consistency between measures, the measures used in analyses were square root transformed.

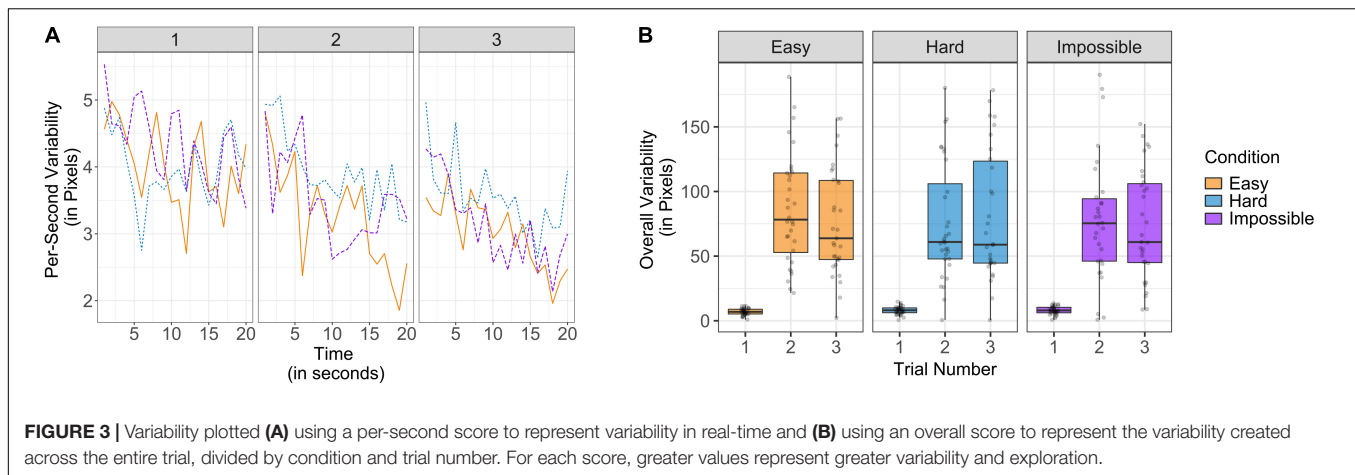
To understand how per-second variability was affected by the effort and success of the adult model, and firsthand experience with failure, a mixed effects model with the same specifications as the previous models was constructed. Participants were entered as random effects and the main effects of condition, trial number, and time within the trial (in seconds) were entered as predictors, the main effects of trial length and motor skill as covariates, as well as the interaction between trial number and time. As before, we expected that the varying success of the adult model between conditions could have an effect on variability. Likewise, we expected that firsthand evidence could have an effect, though we did not have specific hypotheses as to whether per-second variability would decrease, as failure may encourage lesser overall engagement, or increase, as failure may also potentiate exploration. This model revealed only trending effects of time such that per-second variability decreased over time [$t(5479) = -1.84$, $p = 0.07$, $\beta = -0.03$, $SE = 0.02$], and a trending interaction between trial and time [$t(5479) = -1.89$, $p = 0.06$, $\beta = -0.02$, $SE = 0.01$], such that per-second variability decreased to a greater extent on later trials. There were no effects of condition nor trial number (all p 's > 0.24). Likewise, we did not observe an effect of motor skill on per-second variability ($p = 0.35$). Thus, as infants experienced greater firsthand failure,

they exhibited less real-time variability, but we did not find evidence of an effect of the effort and success of the adult model.

On the other hand, to understand how overall variability varied as a function of condition and trial number, a linear mixed effects model was built to predict overall variability. In this case, we hypothesized that overall variability should increase across trials as greater firsthand experience with failure should suggest that previously tested solutions would not succeed. Participants were entered as random effects and the main effects of trial number and condition were entered as predictors. We also entered the main effects of trial length and motor skill into the model as covariates. As before, we used the Easy condition as a baseline, though it is worth noting the pattern of results was the same regardless of baseline. We found a main effect of trial number, such that spatial variability increased across trials [$t(222) = 13.35$, $p < 0.001$, $\beta = 2.71$, $SE = 0.20$], but there were no significant effects of condition (both p 's > 0.43). Therefore, while infants' spatial variability responded to firsthand failure across trials, we did not find evidence that it was also sensitive to the effort and success of the adult model. As in our other models, motor skill did not have a significant effect on overall variability ($p = 0.74$). This analysis of overall variability revealed a markedly different pattern than our analysis of per-second variability. We discuss the potential explanations and implications of these results in the Section "Discussion."

Predicting Individual Differences in Exploration

Our second analytic goal was to understand how individual differences in performance measures predicted infants' exploration, in order to better understand the processes that lead to exploration. The distributions of many of our performance measures exhibited substantial positive skew. While we transformed these variables as necessary to reduce skewness (e.g., trying time, maximum force, overall variability), we additionally employed 20% percentage-bend correlations to increase the robustness of analyses predicting exploration utilizing negative affect, maximum pulling force, help-seeking, and trying time, respectively. Pearson correlations may lack



robustness with this type of data, as small shifts in marginal distributions or outliers can lead to substantial variations in correlation estimates (Wilcox, 1994). Thus, utilizing 20% percentage-bend correlations allowed our analyses to have greater robustness against the skew exhibited in our performance measures. We conceptually treated each performance measure as a predictor of imitative similarity and the square root of overall variability, respectively. However, as we did not have specific hypotheses about the direction of the effects, these analyses were all exploratory and correlational.

We first looked to see how the performance measures related to imitative similarity. Increases in maximum pulling force were related to decreases in imitative similarity ($\rho_{pb} = -0.16$, $p = 0.03$). Thus, infants who utilized greater maximum pulling force tended to diverge more from imitation. However, there were no other trending or significant relationships observed between imitative similarity and the performance measures (all p 's > 0.28 ; see Figure 4). We next performed individual difference analyses of spatial variability (see Figure 5). Regarding force, we found that increases in maximum pulling force were associated with greater overall variability ($\rho_{pb} = 0.22$, $p = 0.003$). Therefore, infants who utilized greater maximum pulling force tended to generate greater spatial variability. We also found that increases in trying time were associated with lower spatial variability ($\rho_{pb} = -0.27$, $p < 0.001$). Lastly, there was a trending relationship with affect and variability. We found that increased negative affect tended to be associated with greater overall variability ($\rho_{pb} = 0.13$, $p = 0.08$). Thus, infants who were more frustrated may have generated greater spatial variability. Finally, we did not find evidence of a relationship between help-seeking and overall variability ($\rho_{pb} = 0.11$, $p = 0.13$).

Predicting Differences in Expectations of Task Success

Finally, we conducted 20% percentage-bend correlations to investigate the relationship between expectations of task success (i.e., the number of hints infants needed on the recovery trial) and imitative similarity and spatial variability, respectively. We hypothesized that infants who explored more in preceding test

trials would require less support during the new iteration of the task, as each new strategy employed would present a new opportunity for success. Since there was only one measure of hints required during the recovery trial for each participant, we averaged the standard deviation of spatial displacement across the three trials, as well as the additive imitative similarity scores, for analyses. There was not a significant relationship between hints required during the recovery trial and imitative similarity ($p = 0.38$). However, hints required and overall variability were moderately, negatively correlated ($\rho_{pb} = -0.37$, $p < 0.001$) such that infants who had higher average overall variability during the test trials required fewer hints during the recovery trials. Importantly, our measures of spatial variability and hints required during the recovery trial were independently collected. Since spatial variability was measured prior to the recovery trial, it seems that spatial variability while problem-solving predicted the number of hints required in the recovery trial.

DISCUSSION

Insights Gained About Infants' Problem-Solving Strategies

This paper's primary conceptual objective was to investigate the influence of the effort and success of an adult model, and firsthand experience with failure on infants' problem-solving approaches by quantifying the extent to which these attempts deviated from modeled solutions. To this end, we considered multifaceted components of infants' problem-solving approaches by applying DLC to generate objective, high-quality data: imitative similarity and spatial variability. Our findings revealed that these exploratory facets of problem-solving were relatively immune to social input (i.e., the effort and success of an adult model) but responded to firsthand failure across and within trials. Thus, although imitative similarity and spatial variability were influenced by some of the same factors that influence the time spent problem-solving (see Lucca et al., 2020), focusing on these new measures of exploration yielded new information about the nature of infants' problem-solving approaches.

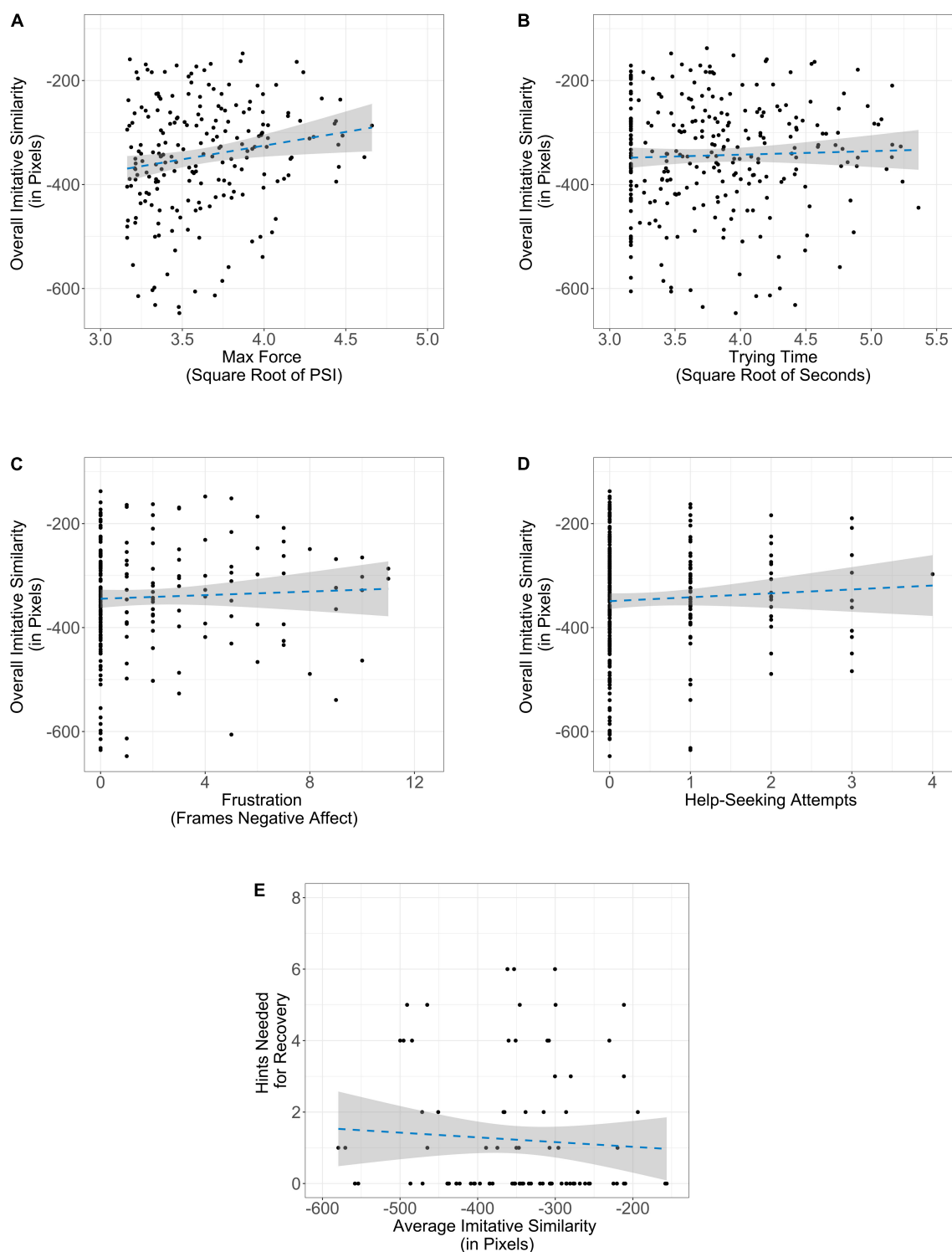


FIGURE 4 | Relationship between performance measures and average imitative similarity: **(A)** maximum pulling force, **(B)** trying time, **(C)** negative affect, **(D)** help-seeking, and **(E)** hints during recovery. The shaded region along the line of best fit represents standard error.

By investigating imitative similarity, we were able to assess the extent to which infants' pulling behaviors deviated from the modeled solution. This process revealed that infants' pulling

behaviors were significantly different from the experimenter model they observed. Thus, when given an unsolvable task, infants as a group generated solutions which were unique from

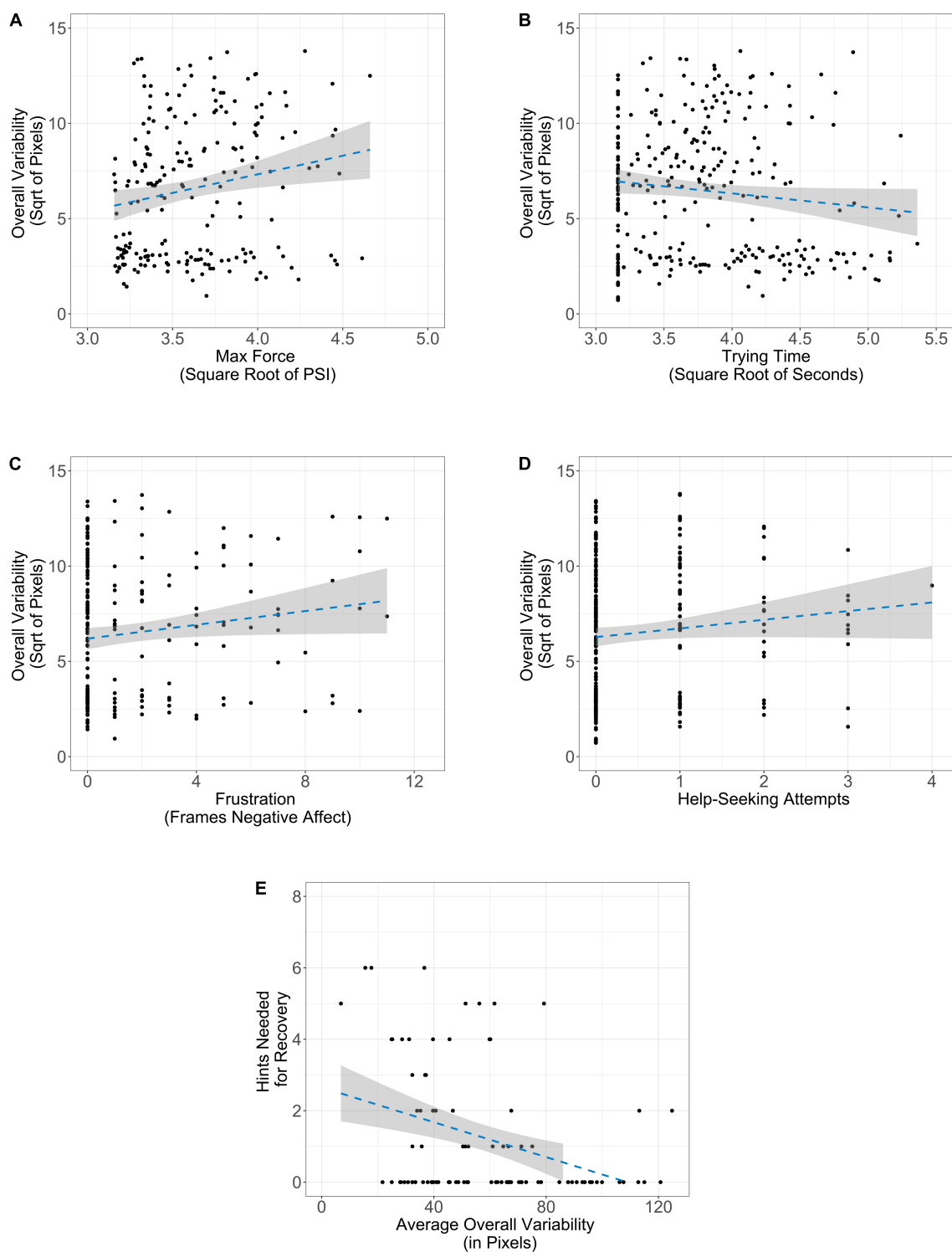


FIGURE 5 | Relationship between performance measures and overall spatial variability: **(A)** maximum pulling force, **(B)** trying time, **(C)** negative affect, **(D)** help-seeking, and **(E)** hints during recovery. The shaded region along the line of best fit represents standard error.

the experimenter by testing new locations. Imitative similarity also varied across time; within trials, infants' approaches tended to become slightly more imitative as they experienced greater

failure, except on later trials, wherein their approaches became less imitative over time. In this task, exploration is useful in that the experimenter's solution demonstrably fails when attempted.

In early trials, infants may test different solutions then converge toward the experimenter's solution once they experience failure with exploration. On the other hand, on later trials, infants may increase exploration after repeated experience failing using the modeled solution. These results suggest that infants' problem-solving approaches respond to firsthand experience with failure.

Likewise, we were able to investigate the variety exhibited in infants' problem-solving approaches by measuring variability, rather than just overall similarity. Our spatial variability findings differed depending on the timescale utilized for analysis. While per-second variability decreased both within and between trials, overall variability actually *increased* between trials. These results can be interpreted as complementary, as standard deviation is highly dependent upon the mean of a given time interval. As such, if infants spent several seconds testing a given location or strategy (i.e., high similarity to the mean of each second), but also tested several disparate locations across the entire trial (i.e., lower similarity to the overall mean), they would demonstrate low per-second variability but high overall variability. Thus, the explanation most consistent with our collection of findings may be that infants persisted for longer in each location they tested as they experienced greater firsthand failure but tested a wider variety of locations throughout the entirety of the trials. Subsequently, spatial variability during test trials predicted recovery trial performance. Infants who produced greater variability during test trials received fewer hints during recovery, requiring less support in the new iteration of the task. Thus, it seems that spatial variability predicted support needed during later recovery trials, suggesting that children who explored more had higher expectations of task success.

Of course, alternate explanations could exist for these results. Most concerningly, our measures of imitative similarity and variability could merely reflect incoordination. However, we find this interpretation quite unlikely because we controlled for motor skill within each of our models. If indeed our measures were merely a reflection of a lack of coordination, we would expect that lower motor skill would be related to each measure. However, we did not observe any significant effects of motor skill on imitative similarity or spatial variability. Likewise, our human coding of unproductive movement was not correlated with imitative similarity or variability. In light of these findings, we find the most consistent explanation is that our results reflect nuanced adjustments to problem-solving approach as infants experienced failure. These adjustments may be deliberate or implicit but are observable in infants' behavior regardless.

While these results display similarities to Lucca et al. (2020), they also indicate departures. In the original study, infants' trying time responded both to the effort and success of the adult model, and to firsthand experience with failure. Thus, both studies indicated an effect of firsthand experience such that infants' problem-solving approaches changed with increased failure, but we did not find evidence for an effect of social input. This is surprising given that similar work also suggests that infants infer appropriate strategies based on social input (e.g., Gweon and Schulz, 2011) and that the demonstrations provided different information about the solution's efficacy. For example, while the Impossible condition cues that the modeled

solution is ineffective, the Hard condition cues that the modeled solution will work eventually. As such, our results point to a potential disassociation between the duration of problem-solving and the approach adopted during problem-solving. It may be the case that exploration represents a more implicit component of problem-solving and responds to firsthand evidence (i.e., failure) but does not become consciously integrated across domains like trying time.

Utility of Applying DeepLabCut

Importantly, this project also sought to illustrate the feasibility and utility of implementing DLC in the analysis of archival data. By identifying a motoric proxy for a cognitive phenomenon (i.e., exploration) we were able to apply computer vision *post hoc* to a previously collected data set to reveal novel insights about problem-solving. This case study provides one example of DLC's application, but the fine-grained data that DLC produces could also be used in more sophisticated computational models and statistical techniques, much like linguistic corpora have been utilized (e.g., Redington et al., 1998; Yang, 2013; Meylan et al., 2017; Bergey et al., 2021). Importantly, DLC is particularly useful when in-person data collection is impossible. DLC can utilize archival data which is an invaluable tool (Gordon et al., 2015), and gives researchers access to high-quality or even rare data (e.g., data from samples which are not Western, Educated, Industrialized, Rich, and Democratic; see Rad et al., 2018; Syed et al., 2018). However, archival data is often collected to answer specific questions and, consequently, the stimulus design may not easily lend itself to new questions. In these cases, DLC provides researchers with open-access tools to answer additional questions that are otherwise very difficult or time-consuming for human coding, extending the lifecycle of existing archival data as in the case of our data. Thus, the advantages of DLC are pertinent for archival research both when in-person data collection is and is not possible.

Implications for Theories of Problem-Solving and Related Phenomena

Classic work demonstrates that young infants have perseverative tendencies, wherein they will continue to apply previously successful solutions to solve problems even when they are no longer effective. Infants' A-not-B task performance classically illustrates this phenomenon: after a 10-s delay, even 12-month-olds demonstrated perseverative errors by continuing to search in the original location an object was hidden instead of its current location (Diamond, 1985). Although perseveration on this particular task diminishes across the second year of life (Lockman and Pick, 1984; McKenzie and Bigelow, 1986; Aguiar and Baillargeon, 1999), perseveration more broadly construed persists into at least the preschool years in various motor-based tasks (Schutte and Spencer, 2002; Mash et al., 2003; Sharon and DeLoache, 2003; DeLoache et al., 2004; Smitsman and Cox, 2008; Schmuckler, 2013). In contrast to these findings, within the context of our study, 18-month-olds demonstrated relative flexibility, testing solutions unique from the adult model, testing

a greater variety of solutions across trials, and varying imitative similarity based on trial number. These findings suggest that perseverative tendencies may vary both with the nature of the task and infants' experiences during the task. Overall, our study juxtaposes previous work on perseveration by showcasing infants' ability to generate productive responses in the face of failure.

Tendencies toward imitation versus exploration can also be understood through the explore-exploit tradeoff, a common framework describing the inherent tension between exploiting known solutions for rewards and taking time to explore better solutions (Mehlhorn et al., 2015). Within this framework, imitation can be understood as exploitation, as infants can conserve mental resources while gaining the benefits of a known solution. Conversely, exploration may produce better solutions but may come at the expense of efficiency, as generating new solutions requires trial-and-error. Longitudinal comparisons have revealed that children tend to explore to a greater extent than adults, choosing to gather information rather than rely on exploiting known effects, with this tendency reducing over development (Gopnik et al., 2015, 2017; Sumner et al., 2019; Gopnik, 2020). However, this research has generally been done with children who are preschool-age or older, due to the cognitive demands of the tasks employed. Our findings demonstrate a ready tendency to explore novel solutions in infants, suggesting that this tendency may be present from an even younger age. Future work could adopt similar paradigms to allow for a full developmental comparison beginning in infancy.

Previous work from this perspective has also differentiated exploration into two subsections: directed and random exploration (Meder et al., 2021; Wilson et al., 2021). Whereas directed exploration serves to sample the areas of greatest uncertainty in a problem space, random exploration simply generates variability. Importantly, both forms of exploration are posed as adaptive, as directed exploration allows for the inspection of features which are likely to produce rewards, but random exploration allows for the discovery of less obvious features which may also be useful. Critically, our methodology did not differentiate between directed and random exploration, as deviation and variability could represent both random, implicit micro-explorations and qualitatively distinct strategies. Thus, future work may adopt methods that elucidate this distinction.

Our findings regarding exploration may also be suited to a larger literature characterizing children's intuitions about effort exertion and problem solving as fundamentally rational. Effort is costly, requiring metabolic resources and creating inherent opportunity costs. As such, children as young as 6 months old expect others to utilize the most efficient paths possible to obtain their goals (Brandone and Wellman, 2009; Scott and Baillargeon, 2013; Skerry et al., 2013; Liu and Spelke, 2017). The naïve utility calculus integrates these intuitions into a framework explaining how children take advantage of the utility of others' actions to infer a wide variety of information including desires, preferences, and prosocial tendencies (Jara-Ettinger et al., 2015a,b, 2016). Recent research has begun to elaborate how children also display effort efficiency in their own actions (Leonard et al., 2020; Lucca et al., 2020; Rett and Walker, 2020). The results presented here are consistent with this larger framework, demonstrating that infants

engaged in several exploratory behaviors which increased the utility of their actions. Infants did not merely copy the approach of the experimenter when they did not experience success, but rather deviated from the demonstration. As infants were confronted with their own failure, they also generally increased their exploration, increasing their deviation from imitation and trying a greater number of exploratory strategies. In other words, infants' approaches responded to information about the productivity of imitation, as well as the productivity of each strategy that they employed. Further, greater exploration related to greater expectations of task success. As infants tested new methods, their expectations of success may have been buffered through failure as there is a possibility that each untested strategy could lead to success. While these results are still speculative, they suggest that infants may engage with problems in nuanced ways to maximize probabilities of success rather than merely giving up or perseverating.

Limitations

The primary limitation of this study is the correlational nature of our data. As with any archival research, if the questions under investigation pertain to variables that were not directly manipulated in the original study, researchers are limited in making causal assertions about their data. In the case of our study, we were only able to make inferential claims about variables which were collected or manipulated independently (i.e., adult modeling, trial number, and recovery trial performance) but further work will be required to make definitive conclusions about the relations observed between other variables, particularly the role of exploration within a larger motivational framework as described in our individual difference analyses. However, correlational work serves as an important exploratory space for generating new research questions and as such, the efficiency of DLC makes it an ideal option for researchers who are endeavoring to test the feasibility or theoretical validity of a research question before investing the time designing an appropriate paradigm, collecting data, and processing data (either in-person, or online).

FUTURE DIRECTIONS AND CONCLUSION

This work provides rich theoretical grounds for future research through its correlational findings. In addition to the directions identified above, we would recommend a further investigation of other facets of exploration and the developmental trajectories of exploratory tendencies. Here, we considered two potential facets of exploration, demonstrating that these two facets responded to firsthand experience with failure. Future work should also consider which other facets may comprise a full constellation of exploratory problem-solving behavior beyond imitative similarity and variability. Likewise, if the measures elaborated in this paper reflect exploration, they raise further questions about the developmental trajectory of these abilities (Muentener et al., 2018). Finally, our correlational analyses of individual differences in spatial variability and other facets of performance raise productive possibilities for

empirical investigation and replication. Exploration may be one component of an interplay between failure and problem-solving. As such, it may explain divergence between learners, in which failure results in divestment for some but growth and learning for others. This interpretation is corroborated by our recovery results, which suggest that greater exploration during problem-solving leads to greater expectations of task success. Perhaps encouraging children who do not naturally produce large exploratory variability to explore will buffer motivational losses.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. These data can be found here: <https://osf.io/sydzq/>.

ETHICS STATEMENT

Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

HS, MR, and JS made substantial contributions to the conception and design of the work. HS implemented DeepLabCut and

coordinated data processing. MR performed all data analyses. MR and JS provided critical oversight and feedback of the work. HS and MR wrote the manuscript. JS provided critical feedback on the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Funding for this research came from the Canadian Foundation for Innovation and the Ontario Research fund to JS (John R. Evans Leaders Fund, Award #38740).

ACKNOWLEDGMENTS

We would like to thank Kelsey Lucca and Rachel Horton for their work that has made this project possible, as well as Ece Yucer and Grace Zheng for their coding. We would also like to thank the families and infants that participated in this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.705108/full#supplementary-material>

REFERENCES

- Adolph, K. E., Eppler, M. A., Marin, L., Weise, I. B., and Wechsler Clearfield, M. (2000). Exploration in the service of prospective control. *Infant Behav. Dev.* 23, 441–460. doi: 10.1016/S0163-6383(01)00052-2
- Aguiar, A., and Baillargeon, R. (1999). Perseveration and problem solving in infancy. *Adv. Child Dev. Behav.* 27, 135–180. doi: 10.1016/S0065-2407(08)60138-X
- Babik, I., Cunha, A. B., Ross, S. M., Logan, S. W., Galloway, J. C., and Lobo, M. A. (2018). Means-end problem solving in infancy: development, emergence of intentionality, and transfer of knowledge. *Dev. Psychobiol.* 61, 191–202. doi: 10.1002/dev.21798
- Barrett, J. M., Raineri Tapies, M. G., and Shepherd, G. M. G. (2020). Manual dexterity of mice during food-handling involves the thumb and a set of fast basic movements. *PLoS One* 15:e0226774. doi: 10.1371/journal.pone.0226774
- Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development*, 3rd Edn. San Antonio, TX: Harcourt Assessment.
- Berger, S. E., and Adolph, K. E. (2003). Infants use handrails as tools in a locomotor task. *Dev. Psychol.* 39, 594–605. doi: 10.1037/0012-1649.39.3.594
- Bergey, C., Marshall, Z., DeDeo, S., and Yurovsky, D. (2021). “Learning communicative acts in children’s conversations: a Hidden Topic Markov Model analysis of the CHILDES corpus,” in *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. (Cognitive Science Society).
- Brandone, A. C., and Wellman, H. M. (2009). You can’t always get what you want: infants understand failed goal-directed actions. *Psychol. Sci.* 20, 85–91. doi: 10.1111/j.1467-9280.2008.02246.x
- Brandt, E. E., Sasiharan, Y., Elias, D. O., and Mhatre, N. (2021). Jump takeoff in a small jumping spider. *J. Comp. Physiol. A Neuroethol. Sens. Neural. Behav. Physiol.* 207, 153–164. doi: 10.1007/s00359-021-01473-7
- Bridgers, S., Wang, Y., and Buchsbaum, D. (2019). “Children’s exploration as a window into their causal learning,” in *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. (Cognitive Science Society).
- Brugger, A., Lariviere, L. A., Mumme, D. L., and Bushnell, E. W. (2007). Doing the right thing: infants’ selection of actions to imitate from observed event sequences. *Child Dev.* 78, 806–824. doi: 10.1111/j.1467-8624.2007.01034.x
- Burke, N. (2019). *Aoianalysis: Functions for AOI Analysis. R Package Version 0.1.0*. Available online at: <https://nicoleburke.github.io/aoianalysis/> (accessed October 14, 2021).
- Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2021). OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Machine Intelligence* 43, 172–186. doi: 10.1109/tpami.2019.2929257
- Carr, K., Kendal, R. L., and Flynn, E. G. (2015). Imitate or innovate? Children’s innovation is influenced by the efficacy of observed behaviour. *Cognition* 142, 322–332. doi: 10.1016/j.cognition.2015.05.005
- Claxton, L. J., Keen, R., and McCarty, M. E. (2003). Evidence of motor planning in infant reaching behavior. *Psychol. Sci.* 14, 354–356. doi: 10.1111/1467-9280.24421
- Cronin, N. J. (2021). Using deep neural networks for kinematic analysis: challenges and opportunities. *J. Biomech.* 123:110460. doi: 10.1016/j.jbiomech.2021.110460
- Cronin, N. J., Rantalainen, T., Ahtiainen, J. P., Hynynen, E., and Waller, B. (2019). Markerless 2D kinematic analysis of underwater running: a deep learning approach. *J. Biomech.* 87, 75–82. doi: 10.1016/j.jbiomech.2019.02.021
- DeLoache, J. S., Uttal, D. H., and Rosengren, K. S. (2004). Scale errors offer evidence for a perception-action dissociation early in life. *Science* 304, 1027–1029.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). “ImageNet: a Large-Scale Hierarchical Image Database,” in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. (IEEE), 248–255.
- Diamond, A. (1985). Development of the ability to use recall to guide action, as indicated by infants’ performance on ab. *Child Dev.* 56, 868–883. doi: 10.2307/1130099

- Esseily, R., Nadel, J., and Fagard, J. (2010). Object retrieval through observational learning in 8- to 18-month-old infants. *Infant Behav. Dev.* 33, 695–699. doi: 10.1016/j.infbeh.2010.07.017
- Fagard, J., and Lockman, J. J. (2009). Change in imitation for object manipulation between 10 and 12 months of age. *Dev. Psychobiol.* 52, 90–99. doi: 10.1002/dev.20416
- Fagard, J., Rat-Fischer, L., Esseily, R., Somogyi, E., and O'Regan, J. K. (2016). What does it take for an infant to learn how to use a tool by observation? *Front. Psychol.* 7:267.
- Fagard, J., Rat-Fischer, L., and O'Regan, J. K. (2014). The emergence of use of a rake-like tool: a longitudinal study in human infants. *Front. Psychol.* 5:491. doi: 10.3389/fpsyg.2014.00491
- Fang, C., Zhang, T., Zheng, H., Huang, J., and Cuan, K. (2021). Pose estimation and behavior classification of broiler chickens based on deep neural networks. *Comput. Electronics Agric.* 180:105863. doi: 10.1016/j.compag.2020.105863
- Fiker, R., Kim, L. H., Molina, L. A., Chomiak, T., and Whelan, P. J. (2020). Visual gait lab: a user-friendly approach to gait analysis. *J. Neurosci. Methods* 341:108775. doi: 10.1016/j.jneumeth.2020.108775
- Forys, B. J., Xiao, D., Gupta, P., and Murphy, T. H. (2020). Real-time selective markerless tracking of forepaws of head fixed mice using deep neural networks. *eNeuro* 7:ENEURO.0096-20.2020. doi: 10.1523/ENEURO.0096-20.2020
- Fragaszy, D., Simpson, K., Cummins-sebree, S., and Brakke, K. (2016). Ontogeny of tool use: how do toddlers use hammers? *Dev. Psychobiol.* 58, 759–772. doi: 10.1002/dev.21416
- Fried, N. T., Chamessian, A., Zylka, M. J., and Abdus-Saboor, I. (2020). Improving pain assessment in mice and rats with advanced videography and computational approaches. *Pain* 161, 1420–1424. doi: 10.1097/j.pain.0000000000001843
- Gergely, G., Bekkering, H., and Király, I. (2002). Rational imitation in preverbal infants. *Nature* 415:755. doi: 10.1038/415755a
- Gill, S. V., Adolph, K. E., and Vereijken, B. (2009). Change in action: how infants learn to walk down slopes. *Dev. Sci.* 12, 888–902. doi: 10.1111/j.1467-7687.2009.00828.x
- Goldfield, E. C. (1983). The development of control over complementary systems during the second year. *Infant Behav. Dev.* 6, 257–262. doi: 10.1016/S0163-6383(83)80035-6
- Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosop. Trans. R. Soc. B: Biol. Sci.* 375:20190502. doi: 10.1098/rstb.2019.0502
- Gopnik, A., Griffiths, T. L., and Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Curr. Direct. Psychol. Sci.* 24, 87–92. doi: 10.1177/0963721414556653
- Gopnik, A., O'grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., et al. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proc. Natl. Acad. Sci. U.S.A.* 114, 7892–7899. doi: 10.1073/pnas.1700811114
- Gordon, A. S., Millman, D. S., Steiger, L., Adolph, K. E., and Gilmore, R. O. (2015). Researcher–Library collaborations: data repositories as a service for researchers. *J. Librarianship Scholarly Commun.* 3:1238. doi: 10.7710/2162-3309.1238
- Gottwald, J. M., De Bortoli Vizioli, A., Lindskog, M., Nyström, P., Ekberg, T. L., von Hofsten, C., et al. (2017). Infants prospectively control reaching based on the difficulty of future actions: to what extent can infants' multiple-step actions be explained by Fitts' law? *Dev. Psychol.* 53, 4–12. doi: 10.1037/dev0000212
- Gottwald, J. M., and Gredebäck, G. (2015). Infants' prospective control during object manipulation in an uncertain environment. *Exp. Brain Res.* 233, 2383–2390. doi: 10.1007/s00221-015-4308-7
- Gottwald, J. M., Gredebäck, G., and Lindskog, M. (2019). Two-step actions in infancy—the TWAIn model. *Exp. Brain Res.* 237, 2495–2503. doi: 10.1007/s00221-019-05604-0
- Gweon, H., and Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science* 332:1524. doi: 10.1126/science.1204493
- Gweon, H., and Schulz, L. E. (2008). “Stretching to learn: ambiguous evidence and variability in preschoolers,” in *Proceedings of the 30th annual meeting of the Cognitive Science Society*. (Cognitive Science Society), 570–574.
- Haberfehlner, H., Buizer, A. I., Stolk, K. L., van de Ven, S. S., Aleo, I., Bonouvrié, L. A., et al. (2020). Automatic video tracking using deep learning in dyskinetic cerebral palsy. *Gait Posture* 81, 132–133. doi: 10.1016/j.gaitpost.2020.07.100
- Hauf, P., Elsner, B., and Aschersleben, G. (2004). The role of action effects in infants' action control. *Psychol. Res.* 68, 115–125. doi: 10.1007/s00426-003-0149-2
- Ho, C. L. A., Zimmermann, R., Flórez Weidinger, J. D., Prsa, M., Schottdorf, M., Merlin, S., et al. (2021). Orientation preference maps in *microcebus murinus* reveal size-invariant design principles in primate visual cortex. *Curr. Biol.* 31, 733.e7–741.e7. doi: 10.1016/j.cub.2020.11.027
- Huang, Z.-J., He, X.-X., Wang, F.-J., and Shen, Q. (2021). A real-time multi-stage architecture for pose estimation of Zebrafish head with convolutional neural networks. *J. Comput. Sci. Technol.* 36, 434–444. doi: 10.1007/s11390-021-9599-5
- Ingvarsdóttir, K. O., and Balkenius, C. (2020). The visual perception of material properties affects motor planning in prehension: an analysis of temporal and spatial components of lifting cups. *Front. Psychol.* 11:215. doi: 10.3389/fpsyg.2020.00215
- Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. (2016). The naïve utility calculus: computational principles underlying commonsense psychology. *Trends Cogn. Sci.* 20, 589–604. doi: 10.1016/j.tics.2016.05.011
- Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., and Schulz, L. E. (2015a). Children's understanding of the costs and rewards underlying rational action. *Cognition* 140, 14–23. doi: 10.1016/j.cognition.2015.03.006
- Jara-Ettinger, J., Tenenbaum, J. B., and Schulz, L. E. (2015b). Not so innocent: toddlers' inferences about costs and culpability. *Psychol. Sci.* 26, 633–640. doi: 10.1177/0956797615572806
- Jung, W. P., Kahrs, B. A., and Lockman, J. J. (2015). Manual action, fitting, and spatial planning: relating objects by young children. *Cognition* 134, 128–139. doi: 10.1016/j.cognition.2014.09.004
- Kahrs, B. A., Jung, W. P., and Lockman, J. J. (2013). Motor origins of tool use. *Child Dev.* 84, 810–816. doi: 10.1111/cdev.12000
- Keen, R. (2011). The development of problem solving in young children: a critical cognitive skill. *Annu. Rev. Psychol.* 62, 1–21. doi: 10.1146/annurev.psych.031809.130730
- Kim, D., Jeong, Y.-C., Park, C., Shin, A., Min, K. W., Jo, S., et al. (2020). Interactive virtual objects attract attention and induce exploratory behaviours in rats. *Behav. Brain Res.* 392:112737. doi: 10.1016/j.bbr.2020.11.2737
- Kominsky, J. F. (2019). PyHab: open-source real time infant gaze coding and stimulus presentation software. *Infant Behav. Dev.* 54, 114–119. doi: 10.1016/j.infbeh.2018.11.006
- Leonard, J. A., Sandler, J., Nerenberg, A., Rubio, A., Schulz, L. E., and Mackey, A. P. (2020). “Preschoolers are sensitive to their performance over time,” in *Proceedings of the 42st Annual Conference of the Cognitive Science Society*. (Cognitive Science Society).
- Liu, H., Reibman, A. R., and Boerman, J. P. (2020). Video analytic system for detecting cow structure. *Comput. Electronics Agric.* 178:105761. doi: 10.1016/j.compag.2020.105761
- Liu, S., and Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition* 160, 35–42. doi: 10.1016/j.cognition.2016.12.007
- Lockman, J. J. (1984). The development of detour ability during infancy. *Child Dev.* 55, 482–491. doi: 10.2307/1129959
- Lockman, J. J. (2000). A perception-action perspective on tool use development. *Child Dev.* 71, 137–144. doi: 10.1111/1467-8624.00127
- Lockman, J. J., and Adams, C. D. (2001). Going around transparent and grid-like barriers: detour ability as a perception-action skill. *Dev. Sci.* 4, 463–471. doi: 10.1111/1467-7687.00188
- Lockman, J. J., and Pick, H. L. (1984). “Problems of scale in spatial development,” in *Origins of Cognitive Skills*, ed. C. Sophian (Hillsdale, NJ: Lawrence Erlbaum Associates), 3–26.
- López Pérez, D., Stryjek, R., and Rączaszek-Leonardi, J. (2021). Recurrence quantification analysis in the study of online coordination in Norway rats (*Rattus norvegicus*). *J. Comparat. Psychol.* 135, 142–149. doi: 10.1037/com0000253
- Loucks, J., and Sommerville, J. A. (2013). Attending to what matters: flexibility in adults' and infants' action perception. *J. Exp. Child Psychol.* 116, 856–872. doi: 10.1016/j.jecp.2013.08.001

- Lucca, K., Horton, R., and Sommerville, J. A. (2020). Infants rationally decide when and how to deploy effort. *Nat. Hum. Behav.* 4, 372–379. doi: 10.1038/s41562-019-0814-0
- Mash, C., Keen, R., and Berthier, N. E. (2003). Visual access and attention in two-year-olds' event reasoning and object search. *Infancy* 4, 371–388.
- Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., et al. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* 21, 1281–1289. doi: 10.1038/s41593-018-0209-y
- Mathis, A., Schneider, S., Lauer, J., and Mathis, M. W. (2020). A primer on motion capture with deep learning: principles, pitfalls, and perspectives. *Neuron* 108, 44–65. doi: 10.1016/j.neuron.2020.09.017
- Mathis, M. W., and Mathis, A. (2020). Deep learning tools for the measurement of animal behavior in neuroscience. *Curr. Opin. Neurobiol.* 60, 1–11. doi: 10.1016/j.conb.2019.10.008
- McCarty, M. E., Clifton, R. K., and Collard, R. R. (1999). Problem solving in infancy: the emergence of an action plan. *Dev. Psychol.* 35, 1091–1101. doi: 10.1037/0012-1649.35.4.1091
- McCarty, M. E., and Keen, R. (2005). Facilitating problem-solving performance among 9- and 12-Month-Old infants. *J. Cogn. Dev.* 6, 209–228. doi: 10.1207/s15327647jcd0602_3
- McKenzie, B. E., and Bigelow, E. (1986). Detour behaviour in young human infants. *Br. J. Dev. Psychol.* 4, 139–148.
- Meder, B., Wu, C. M., Schulz, E., and Ruggeri, A. (2021). Development of directed and random exploration in children. *Dev. Sci.* 24:e13095. doi: 10.1111/desc.13095
- Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., et al. (2015). Unpacking the exploration–exploitation tradeoff: a synthesis of human and animal literatures. *Decision* 2, 191–215. doi: 10.1037/dec0000033
- Meylan, S. C., Frank, M. C., Roy, B. C., and Levy, R. (2017). The emergence of an abstract grammatical category in children's early speech. *Psychol. Sci.* 28, 181–192.
- Muentener, P., Herrig, E., and Schulz, L. (2018). The efficiency of infants' exploratory play is related to longer-term cognitive development. *Front. Psychol.* 9:635. doi: 10.3389/fpsyg.2018.00635
- Mundorf, A., Matsui, H., Ocklenburg, S., and Freund, N. (2020). Asymmetry of turning behavior in rats is modulated by early life stress. *Behav. Brain Res.* 393:112807. doi: 10.1016/j.bbr.2020.112807
- Namba, S., Matsui, H., and Zloteanu, M. (2021). Distinct temporal features of genuine and deliberate facial expressions of surprise. *Sci. Rep.* 11:3362. doi: 10.1038/s41598-021-83077-4
- Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., and Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protocols* 14, 2152–2176. doi: 10.1038/s41596-019-0176-0
- Ossmy, O., Gilmore, R. O., and Adolph, K. E. (2020). “AutoViDev: a computer-vision framework to enhance and accelerate research in human development,” in *Advances in Computer Vision - Proceedings of the 2019 Computer Vision Conference CVC. (Advances in Intelligent Systems and Computing, Vol. 944, eds S. Kapoor and K. Arai (Berlin: Springer Verlag), 147–156.*
- Paxton, A., and Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. *Behav. Res. Methods* 45, 329–343. doi: 10.3758/s13428-012-0249-2
- Provasi, J., Dubon, C. D., and Bloch, H. (2001). Do 9- and 12-month-olds learn means-ends relation by observing? *Infant Behav. Dev.* 24, 195–213. doi: 10.1016/S0163-6383(01)00072-8
- Rad, M. S., Martingano, A. J., and Ginges, J. (2018). Toward a psychology of Homo sapiens: making psychological science more representative of the human population. *Proc. Natl. Acad. Sci. U.S.A.* 115, 11401–11405. doi: 10.1073/pnas.1721165115
- Redington, M., Chater, N., and Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cogn. Sci.* 22, 425–469.
- Reindl, E., and Tennie, C. (2018). Young children fail to generate an additive ratchet effect in an open-ended construction task. *PLoS One* 13:e0197828. doi: 10.1371/journal.pone.0197828
- Rendall, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., et al. (2010). Why copy others? Insights from the social learning strategies tournament. *Science* 328, 208–213. doi: 10.1126/science.1184719
- Repacholi, B. M., Meltzoff, A. N., Hennings, T. M., and Ruba, A. L. (2016). Transfer of social learning across contexts: exploring infants' attribution of trait-like emotions to adults. *Infancy* 21, 785–806. doi: 10.1111/infia.12136
- Rett, A., and Walker, C. M. (2020). “Knowing when to quit: children consider access to solutions when deciding whether to persist,” in *Proceedings of the 42st Annual Conference of the Cognitive Science Society.* (Cognitive Science Society).
- Robin, D. J., Berthier, N. E., and Clifton, R. K. (1996). Infants' predictive reaching for moving objects in the dark. *Dev. Psychol.* 32, 824–835. doi: 10.1037/0012-1649.32.5.824
- Rodriguez, G., Moore, S. J., Neff, R. C., Glass, E. D., Stevenson, T. K., Stinnett, G. S., et al. (2020). Deficits across multiple behavioral domains align with susceptibility to stress in 129S1/SvImJ mice. *Neurobiol. Stress* 13:100262. doi: 10.1016/j.ynstr.2020.100262
- Rook, L., and van Knippenberg, D. (2011). Creativity and imitation: effects of regulatory focus and creative exemplar quality. *Creativity Res. J.* 23, 346–356. doi: 10.1080/10400419.2011.621844
- Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., and Frank, M. C. (2019). Childes-db: a flexible and reproducible interface to the child language data exchange system. *Behav. Res. Methods* 51, 1928–1941. doi: 10.3758/s13428-018-1176-7
- Schmuckler, M. A. (2013). Perseveration in barrier crossing. *J. Exp. Psychol.: Hum. Percept. Perform.* 39, 1100–1123. doi: 10.1037/a0031119
- Schulz, L. E., and Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Dev. Psychol.* 43, 1045–1050. doi: 10.1037/0012-1649.43.4.1045
- Schulz, L. E., Standing, H. R., and Bonawitz, E. B. (2008). Word, thought, and deed: the role of object categories in children's inductive inferences and exploratory play. *Dev. Psychol.* 44, 1266–1276. doi: 10.1037/0012-1649.44.5.1266
- Schutte, A. R., and Spencer, J. P. (2002). Generalizing the dynamic field theory of the A-not-B error beyond infancy: three-year-olds' delay-and experience-dependent location memory biases. *Child Dev.* 73, 377–404.
- Schwier, C., van Maanen, C., Carpenter, M., and Tomasello, M. (2006). Rational imitation in 12-month-old infants. *Infancy* 10, 303–311. doi: 10.1207/s15327078in1003_6
- Scott, R. M., and Baillargeon, R. (2013). Do infants really expect agents to act efficiently? A critical test of the rationality principle. *Psychol. Sci.* 24, 466–474. doi: 10.1177/0956797612457395
- Sharon, T., and DeLoache, J. S. (2003). The role of perseveration in children's symbolic understanding and skill. *Dev. Sci.* 6, 289–296.
- Skerry, A. E., Carey, S. E., and Spelke, E. S. (2013). First-person action experience reveals sensitivity to action efficiency in prereaching infants. *Proc. Natl. Acad. Sci. U.S.A.* 110, 18728–18733. doi: 10.1073/pnas.1312322110
- Smitsman, A. W., and Cox, R. F. (2008). Perseveration in tool use: a window for understanding the dynamics of the action-selection process. *Infancy* 13, 249–269.
- Sommerville, J. A., and Woodward, A. L. (2005a). Infants' sensitivity to the causal features of means-end support sequences in action and perception. *Infancy* 8, 119–145. doi: 10.1207/s15327078in0802_2
- Sommerville, J. A., and Woodward, A. L. (2005b). Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition* 95, 1–30. doi: 10.1016/j.cognition.2003.12.004
- Stahl, A. E., and Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science* 348, 91–94. doi: 10.1126/science.aaa3799
- Sturman, O., von Ziegler, L., Schläppli, C., Akyol, F., Privitera, M., Slominski, D., et al. (2020). Deep learning-based behavioral analysis reaches human accuracy and is capable of outperforming commercial solutions. *Neuropsychopharmacology* 45, 1942–1952. doi: 10.1038/s41386-020-0776-y
- Sumner, E. S., Steyvers, M., and Sarnecka, B. W. (2019). “It's not the treasure, it's the hunt: children are more explorative on an explore/exploit task than adults,” in *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 2891–2897.
- Syed, M., Santos, C., Yoo, H. C., and Juang, L. P. (2018). Invisibility of racial/ethnic minorities in developmental science: implications for research and institutional practices. *Am. Psychol.* 73, 812–826. doi: 10.1037/amp0000294
- Thelen, E., Corbetta, D., and Spencer, J. P. (1996). Development of reaching during the first year: role of movement speed. *J. Exp. Psychol.: Hum. Percept. Perform.* 22, 1059–1076. doi: 10.1037/0096-1523.22.5.1059

- Vonstad, E. K., Su, X., Vereijken, B., Bach, K., and Nilsen, J. H. (2020). Comparison of a deep learning-based pose estimation system to marker-based and kinect systems in exergaming for balance training. *Sensors* 20:6940. doi: 10.3390/s20236940
- Wei, K., and Kording, K. P. (2018). Behavioral tracking gets real. *Nat. Neurosci.* 21, 1146–1147. doi: 10.1038/s41593-018-0215-0
- Whishaw, I. Q., Ghasroddashti, A., Mirza Agha, B., and Mohajerani, M. H. (2020). The temporal choreography of the yo-yo movement of getting spaghetti into the mouth by the head-fixed mouse. *Behav. Brain Res.* 381:112241. doi: 10.1016/j.bbr.2019.112241
- White, M. S., Brancati, R. J., and Lepley, L. K. (2020). Relationship between altered knee kinematics and subchondral bone remodeling in a clinically translational model of ACL injury. *J. Orthopaedic Res. [Online ahead of print]* doi: 10.1002/jor.24943
- Wilcox, R. R. (1994). The percentage-bend correlation coefficient. *Psychometrika* 59, 601–616. doi: 10.1007/BF02294395
- Willatts, P. (1990). "Development of problem-solving strategies in infancy," in *Children's Strategies: Contemporary Views of Cognitive Development*, ed. D. F. Bjorklund (Hove: Psychology Press), 23–66.
- Willatts, P. (1999). Development of means–end behavior in young infants: pulling a support to retrieve a distant object. *Dev. Psychol.* 35, 651–667. doi: 10.1037/0012-1649.35.3.651
- Willatts, P., and Rosie, K. (1988). *Planning by 12-Month-Old Infants*. Paper presented at the British Psychological Society Developmental Section Conference, Harlech, Wales. Harlech.
- Williams, S., Zhao, Z., Hafeez, A., Wong, D. C., Relton, S. D., Fang, H., et al. (2020). The discerning eye of computer vision: can it measure Parkinson's finger tap bradykinesia? *J. Neurol. Sci.* 416:117003. doi: 10.1016/j.jns.2020.117003
- Wilson, R. C., Bonawitz, E., Costa, V. D., and Ebitz, R. B. (2021). Balancing exploration and exploitation with information and randomization. *Curr. Opin. Behav. Sci.* 38, 49–56.
- Wisdom, T. M., and Goldstone, R. L. (2011). Innovation, imitation, and problem solving in a networked group. *Nonlinear Dynamics-Psychol. Life Sci.* 15:229.
- Wu, J. J.-S., Hung, A., Lin, Y.-C., and Chiao, C.-C. (2020). Visual attack on the moving prey by cuttlefish. *Front. Physiol.* 11:648. doi: 10.3389/fphys.2020.00648
- Yang, C. (2013). Ontogeny and phylogeny of language. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6324–6327. doi: 10.1073/pnas.1216803110
- Zmyj, N., Daum, M. M., and Aschersleben, G. (2009). The development of rational imitation in 9- and 12-month-old infants. *Infancy* 14, 131–141. doi: 10.1080/15250000802569884

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Solby, Radovanovic and Sommerville. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Creating a Corpus of Multilingual Parent-Child Speech Remotely: Lessons Learned in a Large-Scale Onscreen Picturebook Sharing Task

Fei Ting Woon^{1*}, Eshwaaree C. Yogarajah¹, Seraphina Fong¹,
Nur Sakinah Mohd Salleh¹, Shamala Sundaray¹ and Suzy J. Styles^{1,2,3}

¹ Psychology, Nanyang Technological University, Singapore, Singapore, ² Centre for Research and Development in Learning, Nanyang Technological University, Singapore, Singapore, ³ Agency for Science, Technology and Research, Singapore Institute for Clinical Sciences, Singapore, Singapore

OPEN ACCESS

Edited by:

Sho Tsuji,
The University of Tokyo, Japan

Reviewed by:

Eunkyoung Shin,
Virginia Tech, United States
Hiromichi Hagihara,
The University of Tokyo, Japan

*Correspondence:

Fei Ting Woon
feitingwoon@ntu.edu.sg

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 01 July 2021

Accepted: 19 October 2021

Published: 19 November 2021

Citation:

Woon FT, Yogarajah EC, Fong S, Salleh NSM, Sundaray S and Styles SJ (2021) Creating a Corpus of Multilingual Parent-Child Speech Remotely: Lessons Learned in a Large-Scale Onscreen Picturebook Sharing Task.
Front. Psychol. 12:734936.
doi: 10.3389/fpsyg.2021.734936

With lockdowns and social distancing measures in place, research teams looking to collect naturalistic parent-child speech interactions have to develop alternatives to in-lab recordings and observational studies with long-stretch recordings. We designed a novel micro-longitudinal study, the Talk Together Study, which allowed us to create a rich corpus of parent-child speech interactions in a fully online environment (N participants = 142, N recordings = 410). In this paper, we discuss the methods we used, and the lessons learned during adapting and running the study. These lessons learned cover nine domains of research design, monitoring and feedback: Recruitment strategies, Surveys and Questionnaires, Video-call scheduling, Speech elicitation tools, Videocall protocols, Participant remuneration strategies, Project monitoring, Participant retention, and Data Quality, and may be used as a primer for teams planning to conduct remote studies in the future.

Keywords: online assessment, multilingualism, corpus creation, parent-child interaction, book sharing

INTRODUCTION

Singapore is a diverse environment for studying language acquisition, with 74.3% of the population reporting literacy in two or more languages (Department of Statistics of Singapore, 2021). In late 2019, our team was preparing for a large-scale project to create a corpus of 500 linguistically diverse home-recordings, and document patterns of language switching, mixing and translanguaging over the first 4 years of life. In the original research plan, a visit to the family home would initiate a series of audio recordings, including high fidelity recording of the parent's voice in their different languages, a parent-child interaction centred on a picture-book narration task, and a day-long ambient speech recording using a baby-worn recording device [e.g., a Language ENvironment Analysis (LENA) device (Gilkerson and Richards, 2008)].

As cases of COVID-19 emerged in Singapore in January 2020, parents began declining invitations to participate in face-to-face procedures. Strict lockdown measures from 8 April, 2020 precluded visits to family homes (Ministry of Health Singapore, 2020). With our corpus-building goals in mind, we designed a novel study that would allow us to develop a large corpus of parent-child speech. The Talk Together Study is a remote micro-longitudinal study in which parent-child

dyads participated in an online book-sharing session at three time-points, separated by at least 1 month. Embedded in the micro-longitudinal design was a Randomised Control Trial (RCT) of an intervention to enhance child-directed talk through “daily tips”. We preregistered the RCT and a target of 150 participants (for further details, see Sundaray et al., 2021). The adapted design allowed us to create a large corpus of parent-child interactions (typically around 20 min in duration) using a novel, remote procedure. This report describes the strategies adopted and lessons learned while running the study.

In recent years, interest in online behavioural studies has been increasing, with conferences and workshops dedicated to online methods (e.g., BeOnline Conference, 2021) and the emergence of user-friendly interfaces for building online experiments (e.g., Gorilla Online Experiment Builder: Anwyl-Irvine et al., 2019) and managing participant recruitment (e.g., Amazon, 2005¹; Prolific, 2014²). Commonly noted advantages of online research include the possibility of recruiting large samples in a short amount of time, as well as reducing physical barriers to participation for a diverse range of participants, and greater access for under-represented populations such as children, and speakers of non-WEIRD languages (Woods et al., 2015; Evershed, 2021; Nation, 2021). The emergence of the COVID-19 has accelerated transitions to online methods for many research groups.

In infancy research, one innovation has been the development of asynchronous browser-based methods that allow a parent to initiate an online study session in their own time, without synchronous involvement from a researcher. Examples include eye-gaze studies using e-Babylab (Lo et al., 2021) and looking-time studies using Lookit (Scott and Schulz, 2017), a platform that has even been used with infants as young as 7 months-of-age (Bochynska and Dillon, 2021). One innovative global collaboration – Manybabies-AtHome – aims to coordinate researchers around the world to develop and run online asynchronous tasks such as preferential looking paradigm and looking-while-listening in home-based tests of infants from diverse backgrounds and nationalities (Zaadnoordijk et al., 2021).

Compared to the efficiency of recruiting and running asynchronous online studies, synchronous online studies requiring a researcher to interact with the participants may have certain limitations, especially regarding manpower. However, synchronous online studies in which a researcher initiates the parent-child interaction, but does not participate, more closely simulate the role of a researcher in a lab-based study, who is able to check equipment function before beginning, and monitoring the study from a concealed location (e.g., behind a two-way mirror, over a live video feed).

In the study reported here, one of the primary goals was to create an audio and transcription corpus of parent-child interactions. While some asynchronous platforms allow participants to record and upload audio recordings

asynchronously (e.g., Discoveries Online: Rhodes et al., 2000), the issue of data storage and transfer remains challenging for researchers working outside of the jurisdictions that govern these platforms (Scott and Schulz, 2017; Zaadnoordijk et al., 2021), and existing online research platforms do not provide high-fidelity in-application audio recording. In addition, in our own team’s pilot tests, the audio quality of unsupervised participant recordings varied greatly due to (1) variable background noise, and (2) variable recording devices and audio drivers, which may compress and degrade audio signals in various ways.

Given the primary goal of the current study was to build up a corpus of parent-child interactions, we developed new synchronous researcher-led online protocols to meet the need for data quality checks at the onset of recording. In this paper, we discuss methods we developed to create a corpus of parent-child speech in a fully online environment using a wordless picture book “Little Orangutan: What a Scary Storm” (Styles, 2020b). We provide concrete recommendations for teams who wish to conduct researcher-led synchronous online studies by video call.

METHODS

Research Setting

The study was conducted fully online. In line with local health restrictions, at the start of the study, the research team were working from home and parents were recruited to participate from their own homes. Internet usage rates in Singapore are extremely high (International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database, 2020), with over 96% of young and middle-aged Singaporeans using the internet, more mobile data subscriptions than residents, and homes are connected by high bandwidth internet (Infocomm Media Development Authority, 2019a,b,c).

Research Design

There were three time-points for data collection in our study – T1 (baseline), T2 (post-intervention), and T3 (post-intervention switch). At each time-point, parents completed a series of online surveys we had designed about their child’s language exposure, understanding, and use (Styles et al., 2021; Woon et al., 2021). After completion of all surveys at each time-point, the parent-child dyad participated in a recorded video call with a member of the Data Collection team.

Eligibility Criteria

Posts on social media invited Singaporean parents with a child between the ages of 8 and 36 months to join the study. Instances where families took some time to enrol in the study post-consent, families with children up to 40 months of age were considered eligible to participate. Non-Singaporean parents who expressed interest were contacted by email and text messages to assess their eligibility. Since the primary targets of the corpus were local patterns of language use, families were deemed eligible if at least one parent had spent a significant time living in Singapore, and had completed most of their education in Singapore.

¹ www.mturk.com

² www.prolific.co

Recruitment Strategies

We preregistered the study with a target sample size of 150 parent-child dyads and used rolling enrolment to replace any participants who dropped out before a preregistered recruitment stop date. Recruitment used organic reach on social media: Our primary strategy was posting announcements about the study on our lab's Facebook page, with lab members sharing to their personal networks; our secondary strategy was to share these posts on local Facebook parenting groups. These peer-to-peer “mummy groups” are popular in Singapore as they allow parents to post questions, share their parenting woes, and seek advice from fellow parents. Each local Facebook group has around 5000 members, and each time we posted to one of these groups we saw increases in recruitment.

Consent and Survey Chain

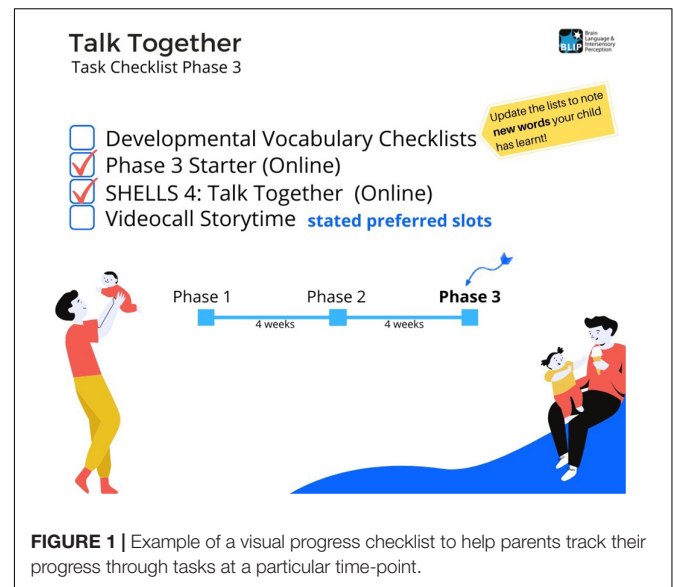
Parents who were interested in taking part in the study were directed to an online consent form in the survey platform, Qualtrics (2005) (Qualtrics, Provo, UT, United States). As informed consent includes an opportunity to ask questions (American Psychological Association, 2017), the online consent page included our lab's contact details. Parents who provided personal details and contact information on the online consent form but did not complete the procedure were contacted by our team, and we answered any questions they had. We also monitored the lab's social media posts and accounts. In all advertising materials, we shared a lab mobile phone number, used to answer queries during the consent process, and throughout the length of the study. The majority of inquiries we received were through asynchronous text-based messaging platforms (WhatsApp, Facebook messenger, or SMS text messages).

The consent form generated a unique random ID for each parent-child dyad. Parents were sent a series of surveys, linked through the identifier, making up a “chain”. Each survey was designed to be short (i.e., 5–20 min duration), so that parents could fit them in around other activities. On completion of each survey, a unique hyperlink was generated to the next survey in the chain. The hyperlink was displayed onscreen, along with a visual checklist of progress through the tasks for each time-point (see **Figure 1**). Survey completion also triggered reminder emails to the parent, and to the Data Collection team for monitoring.

Part of the study involved information about specific vocabulary items across the diverse languages of Singapore. We created an in-house vocabulary checklist for including questions about which words a child “understands” and “understands and also says” (c.f., MacArthur Bates CDI: Fenson et al., 1994). After piloting various formats for the checklist, the parent-preferred format was a PDF with clickable checkboxes, which was returned to the parent at subsequent time-points for updating (i.e., adding new words). Parents in our study were asked to return the completed PDFs by email replies.

Video Call Scheduling and Pre-call Checks

The online survey chain ended with a form for parents to select a video call appointment. Parents were sent a text message



and were phoned by the Data Collection team to confirm the timeslot. Parents who had not completed the preliminary tasks were followed up with reminder phone calls and text messages. Parents who did not complete the required tasks within 8 weeks were removed from the main study.

Speech Elicitation Tools

Book sharing is a common activity for young children, and wordless picture books are a well-known tool for eliciting naturalistic speech (Miller et al., 2006; Heilmann et al., 2015; Chaparro-Moreno et al., 2017). The legality of sharing copyright materials is a well-recognised challenge in medical testing (Feldman and Newman, 2013). Some of the best-known wordless picture books for speech elicitation (e.g., “Frog, where are you?”: Mayer, 2003) are protected by copyright. In the context of a lab-based, or home-visit study, a lab can purchase a small number of physical copies, which would then be re-used by all lab-visiting participants. However, in the transition to online studies, sharing a picture book by pdf or “broadcasting” the pages onscreen may infringe on the copyright of a published work, depending on the legal jurisdiction. To ensure our study materials would have applicability in a variety of geographical and technical contexts, we created an open access wordless picture book “Little Orangutan: What a Scary Storm!” (Styles, 2020b). As an added bonus, this open resource lowers usage barriers for researchers with limited funding.

After surveying a number of wordless picture books, the story was designed to focus on the emotions of an animal (an orangutan), in a familiar scenario (caught in the rain). The book is designed as a printable PDF, suitable for printing as an A4 folded booklet. For the online administration of the Talk Together Study, a screen-sharing version was created showing two-pages per spread. This picture book is an open access resource in our growing collection of open-source materials, the SESAME Research Tools [SESAME: Speech Elicitation for Spectral Analysis in Multilingual Environments, (Styles, 2021)].



FIGURE 2 | Illustration of the Videocall Storytime procedure for recording parent-child interactions. Left: zoom-captured image of a parent describing the onscreen wordless picture book. Right: onscreen images from the little orangutan book “What a Scary Storm!” image reproduced with permission (Yogarrajah et al., 2021).

Video-Call Protocol

During task development, we tested several video call platforms for their stability, familiarity, and effectiveness. At the start of the study in June 2020, most parents in Singapore were familiar with the Zoom platform. The Data Collection team created a script and protocol for the recorded video call (Yogarrajah et al., 2021). A day before each scheduled video call, researchers would send a reminder text to the parent with the link to the password-protected Zoom meeting. In this reminder, parents were asked to pick a quiet spot for the call, and to join the video call using a large-screen device (e.g., tablets, laptops, iPads) so that the pictures would display at an appropriate size.

Unlike in completely asynchronous online studies (Rhodes et al., 2000), video-call studies allow a researcher to perform data quality checks before beginning the recording. First, the researcher confirmed that the shared screen was displaying correctly to the parent, on an appropriately-sized device. Second, the researcher checked that the audio quality was clear enough to enable later transcription. In some cases, parents switched devices, moved to a quieter part of their home, adjusted their home internet settings, or changed devices, to improve the audio fidelity of the parent and child’s voices relative to background noise.

After briefing, the researcher used the screensharing function to display the book, and switched off their camera to reduce distractions. Most parents chose to leave their camera on during the recording (see **Figure 2**). After obtaining verbal consent to record, the recording was started, and the Zoom platform displayed an onscreen notice to the parent. In place of book page-turning, parents were asked to say “next”, for the researcher to display the next slide in the deck, and a “jingling” sound was played.

At the end of the book sharing session, the researcher turned on their camera to conduct a debrief with the parents. Parents were asked whether they would consent to “audio release” in addition to their study consent. “Release forms” are common in creative industries so that a subject of a recording can agree to the specific ways in which their image or voice will be used in the future, and the decision is typically made after the subject

knows what has been recorded (Crawford, 2009). In this case, participants were asked if they would allow for their audio recordings to be released to our research team’s open access repository known as the “Growing Collection of Human Voices” (Styles, 2020a). By granting permission for anonymised audio recordings to be released, parents allow uses beyond the original study (e.g., use in a public presentation, as a stimulus in a different study, or as a training dataset). Parents granted release for 385 out of 410 recordings we conducted (94%). Parents were reminded that they may choose to withdraw from the study at any time, or from the Growing Collection repository up until the anonymised recordings have been publicly archived.

Before ending the call, parents had an opportunity to ask questions, and would sometimes seek information about language development from our Data Collection team. These interactions are one of the intangible benefits of conducting online studies with live interactions.

After conducting the video call, the researcher completed an online log which triggered an automated email to the Intervention team to inform them that the parent was now ready to enter the next stage of the RCT. Recordings were downloaded to secured external hard drives and deleted from the Zoom cloud storage at the earliest possible time.

Participant Remuneration Strategies

Participants were paid a token of appreciation upon completion of the video call at each time-point. We decided on cash payment rather than gift cards or vouchers, for the following ethical reasons: Firstly, many researchers who use vouchers are aware that not all participant tokens are redeemed during their eligibility period, meaning that the remuneration strategy may have hidden inequalities. Secondly, vouchers may be hard to utilise during a global pandemic, if the vendors in question are inaccessible, unable to make deliveries, or go out of business before voucher redemption. Finally, and most importantly, cash payments do not expire, and are maximally fungible. This means that remuneration is fair (all participants receive their remuneration), and equitable (remuneration can be used for anything including groceries, rent, or debt repayments). Cash

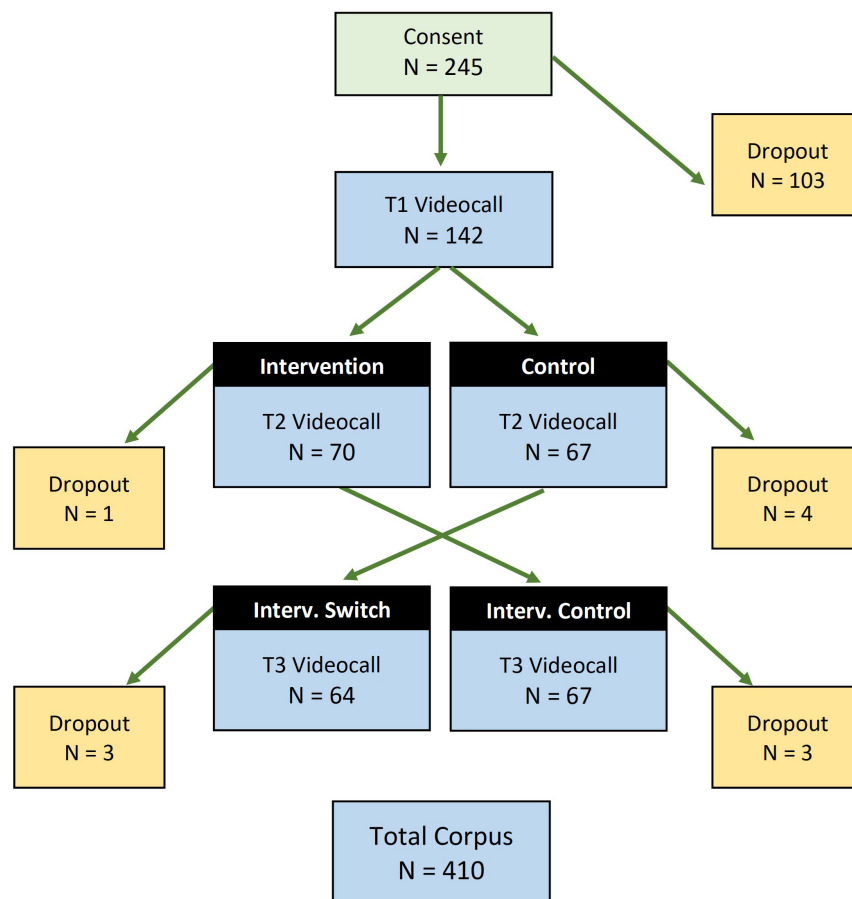


FIGURE 3 | Flowchart of participants and dropouts at different stages in the study. Total corpus $N = 410$ recorded parent-child interaction videos.

payments therefore ensure equitable access to all participants, regardless of income level, and pandemic-related changes to income, housing, or financial status.

In Singapore, electronic payments are ubiquitous, including by registered accounts linked to mobile phone numbers (known as PayNow). The Data Collection team confirmed PayNow details with participating parents at the end of the video call, and filled in an online form while sharing the screen with the parents. The completion of this online form triggered an automated email to our lab's manager who would then proceed with the wireless transfer. Participants were reimbursed upon completion of tasks at each time-point instead of a lump sum transfer at the end of the study.

Despite the ethical advantages of cash reimbursements, we have found that many parents in Singapore do not find cash payments to be strongly motivating. For this reason, in addition to cash reimbursement, we offered lucky draw tokens for each task they completed. We advertised 2 lucky draws, one for participants who completed all tasks at T1 and T2 (10 prizes, approximately 7% of all participants win a prize), and a second draw for participants who continued and completed T3 (5 prizes, 4% of all participants will win a prize). Each lucky draw comprised cash vouchers for the largest local supermarket chain,

a children's book, and a tote bag. While winning the lucky draws mean that the participation remuneration is unequally distributed, our blended model (cash for all, prizes for a few) ensured that all participants were paid an equitable amount prior to the motivational incentive of additional lucky draw rewards.

Project Monitoring

A weekly all-hands project monitoring meeting was held to track progress. Each meeting opened with two check-in questions about unexpected events. Ethics check-in: "Were there any unexpected events that could have ethical consequences for the participants or their data?"; Procedural check-in: "Were there any unexpected events that could have consequences for the way we run our study, or what we can learn from it?". During ethics check-ins, team members discussed situations like what to do when a parent became annoyed or scolded their child during the video call. In these cases, researchers paused the study and reassured parents that the video call was not a formal assessment of their parenting skills or their child's general language abilities, but was a snapshot of language use in Singapore. Ethical training before beginning the study included the concept of *parental consent* coupled with the *child's assent*. In line with best practice in

studies involving minors (Royal College of Paediatrics and Child Health Ethics Advisory Committee, 2000), the Data Collection team reserved the right to stop the video call if they believed a parent was coercing a child to participate without assent.

DELIVERABLES AND PROGRESS

Participant Retention

A total of 245 participants with children aged between 8 and 40 months consented to participate in the remote study. Of these participants, 142 parent-child dyads completed baseline surveys and progressed to the video calls at T1 (see **Figure 3**), thereby beginning the micro-longitudinal study and RCT. Early withdrawals did not complete surveys in the chain asking screening questions (e.g., age of child) and detailed questions about the child's language exposure, family language use, and parenting-related questions. Withdrawals likely indicated that a participant did not have time to participate fully.

Attrition in longitudinal research can be a substantial source of bias in medical and developmental studies (Reinwald et al., 2015; Pan and Zhan, 2020). Although internet-based studies may reduce the overhead required for a family to continue their participation in a longitudinal study, fully asynchronous internet research designs may actually increase dropout rates due to lower participant engagement. The use of video calls with researchers in our Talk Together study combines the ease of access of internet-based methods with the benefits of social interaction that are a part of face-to-face research designs.

Of the 142 parent-child dyads who progressed to the main micro-longitudinal study, we had extremely high participant retention to T2 (96%) and T3 (92% of total), demonstrating that live video call interactions, combined with close project monitoring, can mitigate against attrition in remote studies of this kind. With these low attrition rates across three time-points, we were able to record a total of 410 video calls, making this the largest collection of parent-child speech in Singapore. Transcriptions of the recorded video calls are underway and will be reported separately.

Data Quality

Behavioural studies often have concerns about the fidelity of data collected online relative to lab-based studies, as participants could falsify eligibility criteria, “cheat” to improve their answers, or pay less attention when performing tasks unobserved (Woods et al., 2015; Rodd, 2021). Since our task involves a live video call with parent and child, researchers could monitor engagement, and check eligibility as they would in a lab-visit scenario. In addition, questions in the survey chain about language use in the home also acted as screeners and allowed us to detect ineligible participation (including one family who signed up the same child twice). Since the preregistered study measures will be derived from multilingual transcriptions of the parent-child interactions, our primary data quality concern was whether the audio track in the recording would be sufficiently clear to enable word-level transcriptions. When planning the original study, we conducted a pilot of the parent-child storybook protocol in a

sound attenuating room, using Zoom H4n Pro precision audio recorders. Compared to this small sample of 11 parent-child dyads, the research team noted that the audio quality from the video calls was noisier and of more variable clarity, depending on hardware, internet connectivity, and position in the home. Technology failures like poor Wi-Fi connections, low battery and faulty devices sometimes disrupted data collection, leading to videocall rescheduling. However, since the protocol included an opportunity to change devices or locations, recordings did not proceed until they were judged to be of sufficient audio quality to be transcribed manually.

During periods when social distancing mandates were in place, most parents were working from home, and preschools and childcare services were sporadically unavailable. Unlike in lab-based studies, recordings were occasionally disrupted by other family members, which could disrupt the procedure. When this occurred, the researcher could offer to schedule the call for a later time, or continue the call with requests for the additional people to remain quiet. Data quality checks will continue as our in-lab transcribers monitor task adherence, and create a protocol for further exclusions if necessary.

Reception of the Study by Parents

In general, at the time of the video call, parents gave positive feedback about participating in the study, indicating that the VideoCall Storytime protocol was suitable for parents. Many parents gave positive feedback about the wordless storybook “What a Scary Storm!”, stating that it was fun and interesting, and some even asked whether they could purchase a printed copy of the book.

In our preliminary analyses, almost all parents in the intervention condition reported that the tips provided during the intervention had changed their thinking about language use with their children, and/or their language behaviour with their infants and toddlers (c.f., Amran et al., 2021). Beyond these subjective impacts, transcriptions of the video calls are in progress to find out whether the intervention had a measurable impact on parent's child-directed speech during the video call.

Generalisability

Although the research context provided by Singapore's multilingualism and high technological development is somewhat unusual in global contexts, the lessons learned during the design and delivery of the study have broad applicability for research teams interested in documenting or evaluating parent-child interactions, or other aspects development best captured through live interactions where a trained researcher is present. The procedures could be readily adapted for studies investigating toy play, theory of mind, numerical processing, spatial problem-solving, motor development, and a variety of other developmental milestones. We were able to conduct the study using Zoom, in part due to the prevalence of stable, high-bandwidth home internet connections in Singapore. Although these technological overheads may not be available in some research contexts, lower bandwidth versions of the protocol may be achieved by serving onscreen stimuli by a separate slide-sharing website and recording only audio

TABLE 1 | Summary of lessons learned.

| | |
|----------------------------|---|
| 1 Recruitment strategies | <p>Having an established virtual presence before recruitment (e.g., an active Facebook page) can help to build community trust in your studies.</p> <p>Online parenting groups are a great way to connect with potential participants, using posts and shares rather than paid ads.</p> <p>It is important to monitor social media for questions from potential participants before and during the consent procedure.</p> |
| 2 Consent and survey chain | <p>A dedicated lab handset with mobile data, and lab-based social media accounts allow parents to ask questions and interact using asynchronous text-based messaging during consent, and throughout a study.</p> <p>Breaking up long surveys into smaller sections allows parents to break up the task into manageable chunks, with feedback on progress.</p> <p>Visual checklists help communicate study progress to participants.</p> <p>Automated emails to participants at each stage in the survey chain help them remember what they should do next.</p> <p>Automated emails to the research team allow tracking of task progress.</p> <p>Long surveys with a consistent format may be easier to complete using a clickable PDF, as this reduces “load time” in digital survey environments.</p> <p>Swapping between online surveys and PDF surveys that must be emailed back to the research team may cause some confusion for participants and needs to be monitored with care.</p> |
| 3 Video call scheduling | <p>Reminder messages and calls help busy parents to remember stages in the study process they have missed or forgotten.</p> <p>Pre-call survey completion checks reduce data loss as no video call is conducted without the contextual data required for analysis.</p> |
| 4 Speech elicitation tools | <p>Wordless picture books give a parent-child dyad something to focus on during their interaction. The stimulus is the same for all parents, but the choice of language(s), the level of complexity, and the choice of vocabulary is unconstrained.</p> <p>Open access research tools enhance the variety of contexts in which a tool can be used (including online), and lower barriers to use of the tool for researchers with limited funding.</p> <p>Wordless picture books reduce bias in a parent’s use of particular linguistic varieties, registers and speech styles. Study materials designed for use in multilingual or contact language contexts should be designed for that context, rather than imposing monolingual modes of language use as default.</p> |
| 5 Video call protocols | <p>Synchronous video calls initiated by a researcher allow the research team to optimise the audio and video quality before beginning the recording. Video recordings provide helpful supplementary context for transcribers.</p> <p>Asking a participant to switch on a camera in their own home may be more complex in some contexts than others. For example, some religious communities may need to ensure non-participating members of the household are out of shot, or dressed differently for the duration of the call.</p> <p>Sensitivity to different home situations and flexibility should be built into the protocol, if possible (e.g., allowing a family to participate with their camera off if necessary), as it may facilitate broader participation.</p> <p>Using a single videoconferencing platform with all participants allows the research team to the protocol for that tool.</p> |
| 6 Participant remuneration | <p>Cash (or wireless cash transfer) is the most ethical form of reimbursement as it is equitable and fungible.</p> <p>Lucky draws are inherently inequitable, but can be highly motivating.</p> <p>A combination of cash reimbursement and lucky draw ensures all participants are ethically reimbursed and provides additional motivation.</p> |
| 7 Project monitoring | <p>All-hands check-ins provide an opportunity for ethical monitoring as well as project progress monitoring.</p> <p>All-hands check-ins allow all team members (including junior lab members) to share progress reports and contribute to project development throughout the research process.</p> <p>Building team feedback into the timeline of a study allows valuable opportunities to refine protocols during data collection and for future studies.</p> <p>Including opportunities for parents to give feedback on their impression of the study allows valuable opportunities to refine protocols while a study is still running, or to learn from participant experience.</p> |
| 8 Participant retention | <p>Surveys for parents to complete in their own time have relatively high non-completion rates (a combination of ineligible respondents, and legitimate-but-busy participants). Using chained surveys with visual reminders sent to participants will increase rate of completion.</p> <p>Scheduling and conducting video calls are costly for a research team’s time and labour.</p> <p>Participants who were able to complete time-consuming surveys and a scheduled video call were highly likely to continue at later time-points.</p> <p>Studies that require both live interactions (e.g., video calls) and solo participation (e.g., survey completion) can use surveys as a screener before committing researcher time to live interactions.</p> |
| 9 Data quality | <p>Combining the flexibility of online research with synchronous researcher-led methods allows research staff to optimise data quality before initiating the recording of a research data object</p> <p>Online recording conditions may not be as controlled as in-lab recording conditions, making them more suitable for some types of research measurements (e.g., transcription of speech, coding of behaviour) than others (e.g., precision eye-tracking, fine-grained acoustic analysis). Researchers should pilot their planned procedures to check that they meet the data quality required for their planned analysis before beginning a large-scale study.</p> |

during synchronous online sessions. One drawback of remote digital interactions may be recruitment bias toward those with higher SES and technological skills (Rodd, 2021). However, we

believe these limitations are balanced by the ability for online studies to broaden participation among groups who may not otherwise be able to travel for an in-lab visit. In addition, our

experience during this socially distanced COVID-19 pandemic has suggested this method is viable for reaching out to a variety of participants when physical travel is limited.

Lessons learned during the design, development, and running of the Talk Together Study are summarised in **Table 1**.

CONCLUSION

Over the 12 months of the Talk Together Study, our research team was able to make hundreds of recordings of naturalistic parent-child interactions using a novel procedure called Videocall Storytime. The procedure was contact-free, socially distanced, and possible to run during even the strictest lockdown conditions. Researcher-initiated video calls allowed for data quality checks to precede the onset of recording, and the interactions between parents and researchers provided supplementary motivation to continue in the longitudinal study. The lessons learned during adapting and running the study cover 10 domains of research design, monitoring and feedback: Recruitment strategies, Surveys and Questionnaires, Video call scheduling, Speech elicitation tools, Video call protocol, Participant remuneration, Project monitoring, Participant retention, Parental feedback, and Research team feedback. These lessons may have broad applicability in future research that extends the bounds of research with children beyond the constraints of face-to-face interactions between researchers, children, and their families.

DATA AVAILABILITY STATEMENT

Materials used in this study are publicly archived, and available here: Videocall Storytime Protocol (<https://doi.org/10.21979/N9/0UYKJC>) Language Experiences Overview (<https://doi.org/10.21979/N9/XQUFEW>) SHELLS: Supportive Home Environments for Language Learning – Survey (<https://doi.org/10.21979/N9/RL5UMY>) Talk Together – A 4 week programme of tips to enhance parent-child interactions (<https://doi.org/10.21979/N9/W1D24L>).

REFERENCES

- Amazon (2005). *Mechanical Turk*. Available online at: <https://www.mturk.com> (accessed June 1, 2021).
- American Psychological Association (2017). *Ethical Principles of Psychologists and Code of Conduct*. Available online at: <https://www.apa.org/ethics/code/ethics-code-2017.pdf> (accessed September 22, 2021).
- Amran, S., Sundaray, S., Le, T. A., Woon, F. T., Yogarajah, E. C., and Styles, S. J. (2021). *Engagement in a Remote Intervention: Feedback from a Text-Based Intervention to Enhance Parent-Child Talk*. Society for Research in Child Development Biennial Meeting. Available online at: <https://srcd21biennial.ipostersessions.com/Default.aspx?s=40-43-E3-F7-46-18-23-E1-2E-E1-6E-A1-10-4B-71-FA> (accessed June 1, 2021).
- Anwyl-Irvine, A. L., Massoné, J., Flitton, A., Kirkham, N. Z., and Evershed, J. K. (2019). Gorilla in our midst: an online behavioural experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x
- BeOnline Conference (2021). *BeOnline Conference: The Conference About Online Behavioural Research*. Available online at: <https://beonlineconference.com/> (accessed September 22, 2021).
- Bochynska, A., and Dillon, M. R. (2021). *Bringing Home Baby Euclid: Infants' Basic Shape Discrimination Tested on the Online Platform Lookit*. Available online at: <https://osf.io/vvaw7/> (accessed September 22, 2021).
- Chaparro-Moreno, L. J., Reali, F., and Maldonado-CarreñoCarre, C. (2017). Wordless picture books boost preschoolers' language production during shared reading. *Early Child. Res. Q.* 40, 52–62. doi: 10.1016/j.ecresq.2017.03.001
- Crawford, T. (2009). *Business and Legal Forms for Photographers*. New York, NY: Allworth Press.
- Department of Statistics of Singapore (2021). *Singapore Census of Population 2020, Statistical Release 1: Demographic Characteristics, Education, Language and Religion*. Available online at: https://www.singstat.gov.sg/publications/reference/cop2020/cop2020-sr1/census20_stat_release1 (accessed June 29, 2021).
- Evershed, J. (2021). *How Taking Research Online Can Help Counteract the Replication Crisis*. Available online at: <https://gorilla.sc/online-research-counteracts-replication-crisis/> (accessed September 22, 2021).
- Feldman, R., and Newman, J. (2013). Copyright at the bedside: should we stop the spread? *Stanf. Technol. Law Rev.* 16:623.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Nanyang Technological University IRB Board. Written informed consent to participate in this study was provided by the parent participants for themselves and their children.

AUTHOR CONTRIBUTIONS

FTW, SS, and SJS contributed to the design of the Talk Together Study and prepared the figures. ECY, SF, NSMS, and FTW contributed to the Videocall Storytime protocols. SS and FTW contributed to the project monitoring. FTW and SJS wrote the manuscript with contributions from all other authors. All authors approved the submitted version.

FUNDING

This project was supported by Singapore's National Research Foundation under the Science of Learning grant, 'How do language mixes contribute to effective bilingualism and effective biliteracy in Singapore' (NRF2016-SOL002-011) and Nanyang Technological University under CRADLE@NTU grant (JHU IO 90071537) and NAP Start Up Grant (M4081215.100) awarded to SJS.

ACKNOWLEDGMENTS

The authors would like to thank the following people for their contributions to different elements of the study, including Shaza binte Amran, Tuan Anh Le, Defu Yap, Jinyi Wong, Wai Tung Leung, and Wen Xin Ang.

- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., et al. (1994). In communicative development. *Monogr. Soc. Res. Child Dev.* 59, 174–185.
- Gilkerson, J., and Richards, J. A. (2008). *The Power of Talk (LENA Foundation Technical Report LTR-01-2)*. Available online at: <http://www.lenafoundation.org/TechReport.aspx/PowerOfTalk/LTR-01-2> (accessed September 22, 2021).
- Heilmann, J. J., Rojas, R., Iglesias, A., and Miller, J. F. (2015). Clinical impact of wordless picture storybooks on bilingual narrative language production: a comparison of the “Frog” stories. *Int. J. Lang. Commun. Disord.* 51, 339–345. doi: 10.1111/1460-6984.12201
- Infocomm Media Development Authority (2019a). *Individual Computer Usage (Singapore)*. Available online at: <https://data.gov.sg/dataset/individual-computer-usage> (accessed September 22, 2021).
- Infocomm Media Development Authority (2019b). *Internet Connection at Home by Type*. Available online at: <https://data.gov.sg/dataset/internet-connection-at-home-by-type> (accessed September 22, 2021).
- Infocomm Media Development Authority (2019c). *Mobile Penetration Rate*. Available online at: <https://data.gov.sg/dataset/mobile-penetration-rate> (accessed September 22, 2021).
- International Telecommunication Union (ITU) World Telecommunication/ICT Indicators Database (2020). *Individuals Using the Internet (% of population) – Singapore*. Available online at: <https://data.worldbank.org/indicator/IT.NET.USER.ZS> (accessed September 22, 2021).
- Lo, C. H., Mani, N., Kartushina, N., Mayor, J., and Hermes, J. (2021). e-Babylab: an open-source browser-based tool for unmoderated online developmental studies. *PsyArXiv* [Preprint] doi: 10.31234/OSF.IO/U73SY
- Mayer, M. (2003). *Frog, Where are You*. New York, NY: Penguin Random House.
- Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., and Francis, D. J. (2006). Oral language and reading in bilingual children. *Learn. Disabil. Res. Pract.* 21, 30–43. doi: 10.1177/002205748116300105
- Ministry of Health Singapore (2020). *Circuit Breaker to Minimise Further Spread of COVID-19*. Available online at: <https://www.moh.gov.sg/news-highlights/details/circuit-breaker-to-minimise-further-spread-of-covid-19> (accessed June 20, 2020).
- Nation, K. (2021). *Online Large-Scale Studies with Children Out of the Lab: The Promise and the Challenge*. Available online at: <https://beonlineconference.com/online-large-scale-studies-with-children-out-of-the-lab-the-promise-and-the-challenge/> (accessed September 22, 2021).
- Pan, Y., and Zhan, P. (2020). The impact of sample attrition on longitudinal learning diagnosis: a prolog. *Front. Psychol.* 11:1051. doi: 10.3389/FPSYG.2020.101051
- Prolific (2014). *Prolific*. Available online at: <https://www.prolific.co> (accessed September 22, 2021).
- Qualtrics (2005). *Qualtrics*. Available online at: <https://www.qualtrics.com> (accessed September 22, 2021).
- Reinwald, D. A., Crutzen, R., Elfeddali, I., Schneider, F., Schulz, D. N., Smit, E. S., et al. (2015). Impact of educational level on study attrition and evaluation of web-based computer-tailored interventions: results from seven randomized controlled trials. *J. Med. Internet Res.* 17:e228. doi: 10.2196/JMIR.4941
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2000). *Discoveries Online*. Available online at: <http://discoveriesonline.org/> (accessed September 15, 2021).
- Royal College of Paediatrics, and Child Health Ethics Advisory Committee (2000). Guidelines for the ethical conduct of medical research involving children. *Arch. Dis. Child.* 82, 177–182. doi: 10.1136/ADC.82.2.177
- Rodd, J. (2021). Collecting experimental data online: how to maintain data quality when you can't see your participants. *J. Acoust. Soc. Am.* 149:A81.
- Scott, K., and Schulz, L. (2017). Lookit (Part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_A_00002
- Styles, S. J. (2020b). *Little Orangutan: What a Scary Storm. DR-NTU. V1*. Available online at: <https://doi.org/10.21979/N9/MJMFV> (accessed March 27, 2020).
- Styles, S. J. (2020a). *Growing Collection of Human Voices – Audio Release Form DR-NTU. V1*. Available online at: <https://doi.org/10.21979/N9/I6KXC6> (accessed January 13, 2020).
- Styles, S. J. (2021). *SESAME Research Tools DR-NTU*. Available online at: <https://researchdata.ntu.edu.sg/dataverse/sesame-research-tools> (accessed September 22, 2021).
- Styles, S. J., Woon, F. T., Yogarajah, E. C., and Mohd Salleh, N. S. (2021). *SHELLS: Supportive Home Environments for Language Learning – Survey. DR-NTU. V1*. Available online at: <https://doi.org/10.21979/N9/RL5UMY> (accessed April 8, 2020).
- Sundaray, S., Yap, D., Woon, F. T., Yogarajah, E. C., Amran, S., Fong, S., et al. (2021). *A Remote Intervention to Enhance Child-Directed Speech: An RCT For Infants and Toddlers. Society for Research in Child Development Biennial Meeting*. Available online at: <https://srcd21biennial.ipostersessions.com/Default.aspx?s=3C-B8-DC-45-DA-C4-56-2B-92-25-50-1E-0D-91-20-6F> (accessed June 1, 2021).
- Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., and Spence, C. (2015). Conducting perception research over the internet: a tutorial review. *PeerJ*. doi: 10.7717/PEERJ.1058
- Woon, F. T., Le, T. A., Amran, S., Ang, W. X., and Styles, S. J. (2021). *Language Experiences Overview (LEO). DR-NTU. V1*. Available online at: <https://doi.org/10.21979/N9/XQUFEW> (accessed April 07, 2021).
- Yogarajah, E. C., Woon, F. T., Mohd Salleh, N. S., Amran, S., Fong, S., and Styles, S. J. (2021). *Videocall Storytime. DR-NTU. V1*. Available online at: <https://doi.org/10.21979/N9/0UYKJC> (accessed April 08, 2021).
- Zaadnoordijk, L., Buckler, H., Cusack, R., Tsuji, S., and Bergmann, C. (2021). A global perspective on testing infants online: introducing ManyBabies-AtHome. *PsyArXiv*. doi: 10.31234/OSF.IO/CNWH5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Woon, Yogarajah, Fong, Salleh, Sundaray and Styles. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Remote Testing of the Familiar Word Effect With Non-dialectal and Dialectal German-Learning 1–2-Year-Olds

Bettina Braun^{1*}, Nathalie Czeke², Jasmin Rimpler³, Claus Zinn⁴, Jonas Probst⁵, Bastian Goldlücke⁵, Julia Kretschmer¹ and Katharina Zahner-Ritter^{1,6}

¹ Department of Linguistics, University of Konstanz, Konstanz, Germany, ² School of Education, University of Leeds, Leeds, United Kingdom, ³ Institute of Phonetics and Speech Processing, University of Munich, Munich, Germany, ⁴ Department of Linguistics, University of Tübingen, Tübingen, Germany, ⁵ Department of Computer and Information Science, University of Konstanz, Konstanz, Germany, ⁶ Department of Phonetics, University of Trier, Trier, Germany

OPEN ACCESS

Edited by:

Sho Tsuji,
The University of Tokyo, Japan

Reviewed by:

Camilla Bouchon,
Université Paris-Est Créteil Val
de Marne, France
Marcello Ferro,
Consiglio Nazionale delle Ricerche
(CNR), Italy

*Correspondence:

Bettina Braun
Bettina.braun@uni-konstanz.de

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 25 May 2021

Accepted: 13 October 2021

Published: 02 December 2021

Citation:

Braun B, Czeke N, Rimpler J,
Zinn C, Probst J, Goldlücke B,
Kretschmer J and Zahner-Ritter K
(2021) Remote Testing of the Familiar
Word Effect With Non-dialectal
and Dialectal German-Learning
1–2-Year-Olds.
Front. Psychol. 12:714363.
doi: 10.3389/fpsyg.2021.714363

Variability is pervasive in spoken language, in particular if one is exposed to two varieties of the same language (e.g., the standard variety and a dialect). Unlike in bilingual settings, standard and dialectal forms are often phonologically related, increasing the variability in word forms (e.g., German *Fuß* “foot” is produced as [fu:s] in Standard German and as [fūəs] in the Alemannic dialect). We investigate whether dialectal variability in children’s input affects their ability to recognize words in Standard German, testing non-dialectal vs. dialectal children. Non-dialectal children, who typically grow up in urban areas, mostly hear Standard German forms, and hence encounter little segmental variability in their input. Dialectal children in turn, who typically grow up in rural areas, hear both Standard German and dialectal forms, and are hence exposed to a large amount of variability in their input. We employ the familiar word paradigm for German children aged 12–18 months. Since dialectal children from rural areas are hard to recruit for laboratory studies, we programmed an App that allows all parents to test their children at home. Looking times to familiar vs. non-familiar words were analyzed using a semi-automatic procedure based on neural networks. Our results replicate the familiarity preference for non-dialectal German 12–18-month-old children (longer looking times to familiar words than vs. non-familiar words). Non-dialectal children in the same age range, on the other hand, showed a novelty preference. One explanation for the novelty preference in dialectal children may be more mature linguistic processing, caused by more variability of word forms in the input. This linguistic maturation hypothesis is addressed in Experiment 2, in which we tested older children (18–24-month-olds). These children, who are not exposed to dialectal forms, also showed a novelty preference. Taken together, our findings show that both dialectal and non-dialectal German children recognized the familiar Standard German word forms, but their looking pattern differed as a function of the variability in the input. Frequent exposure to both dialectal and Standard German word forms may hence have affected the nature of (prelexical and/or) lexical representations, leading to more mature processing capacities.

Keywords: familiar word effect, remote testing, iPad App, word representation, children, German, regional variation, dialect

INTRODUCTION

Testing children's word recognition has become an important cornerstone in developing models of lexical representation during the first two years of life. Developmental psychologists have so far paid little attention to how long-term exposure to more than one variety of a language affects children's word recognition abilities. Hence, the nature of early lexical representations in children who grow up with two varieties of a language (e.g., Standard German and a dialectal variant, henceforth "dialectal" children) remains largely unclear. The main aim of the present study is to compare the recognition of Standard German word forms in dialectal and non-dialectal German children. We present a method to reach these dialectal children, using an App for iPads for remote testing of word form recognition.

We tested children's word form recognition using the familiar word paradigm. In this paradigm, children from around 11 months onward have been shown to attend longer to familiar word lists than to unfamiliar or nonce-word lists, hence showing a preference for known words (familiarity preference), which is taken to reflect successful word form recognition (e.g., Hallé and Boysson-Bardies, 1994; Vihman et al., 2004; Carbajal et al., 2021 for a meta study). Children are commonly tested in the lab in a head-turn preference paradigm, HPP (Hallé and Boysson-Bardies, 1994; Vihman et al., 2004; van Heugten and Johnson, 2014) or a visual-fixation paradigm (Best et al., 2009), both of which employ child-controlled stimulus presentation. We chose the familiar word paradigm for two reasons: First, it focuses on the processing of word forms, which may differ for dialectal children who grow up with two varieties of the same language (Standard German and a dialect). Second, the familiar word paradigm is robust (Carbajal et al., 2021), which makes it suitable for replication outside the lab using an App in a more natural but potentially also more distracting environment.

In the present paper, we study whether exposure to a dialect in addition to the Standard affects German-learning children's word recognition abilities. In Experiment 1, we compare two groups of children: (a) 12–18-month-olds who grow up with Standard German only ("non-dialectal children") and (b) 12–18-month-olds who grow up with Standard German and an additional German variety ("dialectal children"). Both groups are tested outside the lab using an experiment-controlled visual fixation procedure implemented in an App. In Experiment 2, we included older non-dialectal children (18–24 months of age) to test one hypothesis that may explain the different patterns of results for dialectal and non-dialectal children in Experiment 1. In the following, we will give a brief overview of dialectal variation in Germany and focus on the coding of dialectal input in more detail (see section "Dialectal Variation in Germany and the Coding of Dialectal Input"), before we move on to review the literature on early word form recognition (see section "Word Form Recognition in Light of Dialectal Exposure").

BACKGROUND

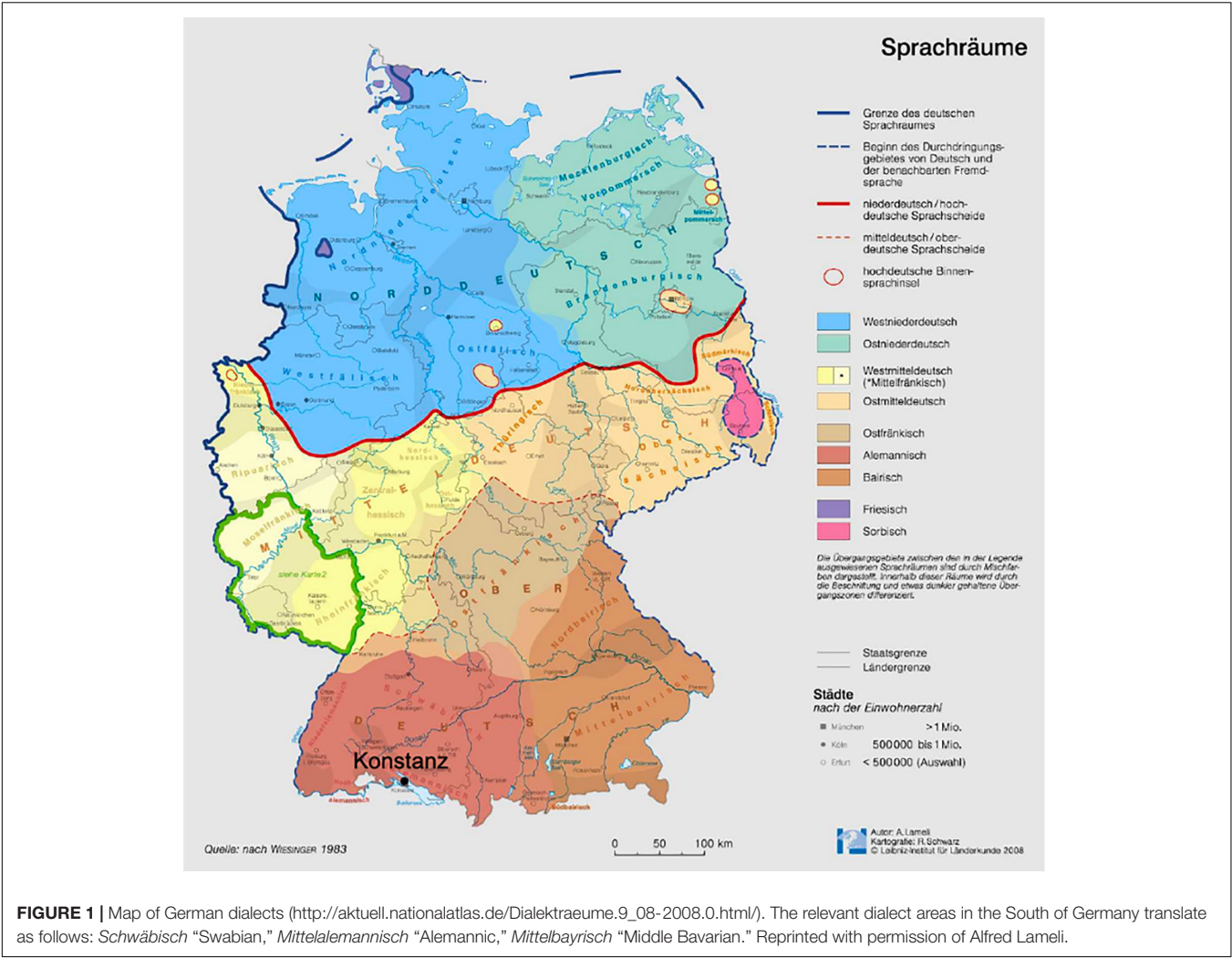
Dialectal Variation in Germany and the Coding of Dialectal Input

Germany is historically divided into different dialectal areas, see **Figure 1**. In their original form, these dialects are difficult to decode for outsiders because they do not only differ in phonology and phonetics, but also in morphology and syntax (Bayer, 1984; Siebenhaar and Wyler, 1997; Brandner and Saltzman, 2009; Grewendorf and Weiß, 2014). We focus on the dialectal areas in Southern Germany (Alemannic and Bavarian, red and orange in **Figure 1**), as most of the dialectal children tested in our study grow up in these regions.

In addition to grammatical and morphological differences, there is a large range of phonological and phonetic differences between dialectal word forms in Alemannic and Bavarian as compared to Standard German forms (Siebenhaar and Wyler, 1997; Munske, 2015). A comprehensive introduction to dialectal phonology is beyond the scope of this paper; we will focus on main differences here (see **Table 1**, for examples). Consonantal differences include lenition of plosives (1), place of articulation of the fricative /s/ (2) in Alemannic, vocalization of coda-liquids (3) in Bavarian, devoicing of word-initial [z] (4), vocalic differences include schwa-elision (5), and diphthongization (6, 7), see **Table 1**.

There is substantial variation in the (proportion of) usage of dialectal forms in Southern Germany (Schwarz, 2014). In more rural areas of Germany, dialect is frequently used for daily communication between locals (Schwarz, 2014). Standard German, the variety present on national TV and in schools, in turn, is spoken when reading to children, in audio books, on radio and TV, and in more formal situations (at the pharmacy, the doctor's etc.). Since Standard German is used in the educational context (school, university), parents may have an incentive to introduce this variety to their children from early on. In any case, Standard German word forms regularly occur in addition to dialectal word forms (around 50% of spontaneous speech tokens in a 2,000 word corpus are dialectal forms in Schwarz, 2014, p. 77), and one and the same caregiver may even switch between dialectal and Standard German forms. The usage of dialectal forms is gradient and a higher proportion of dialectal word forms increases the perceived strength of a speaker's dialect. In contrast, in more urban areas, dialect is (and has increasingly become) less frequent (Schwarz, 2014), probably because the population is more heterogenous, making Standard German the most comprehensible style. Most of our non-dialectal children came from Konstanz, a university city of around 85,000 inhabitants at the lake of Konstanz (see **Figure 1**). The proportion of dialectal forms is substantially smaller in Konstanz compared to more rural areas (Schwarz, 2014, p. 77), with Standard German prevailing in most contexts.

Children growing up in Southern Germany hence differ in the amount of exposure to dialectal word forms they receive. In a recent study on the phonological variability in infant-directed speech, Zahner et al. (2021) showed that around one third of the word forms of dialectal parents contained a dialectal



feature (i.e., a segmental deviation from the Standard form), while two thirds of the word forms were (congruent with) Standard German. Children growing up in Konstanz, on the other hand, were only exposed to around 5% of dialectal forms, with a huge majority of word forms being Standard German. These proportions reported in Zahner et al. (2021) were constant across different recording settings (naturalistic home recordings

vs. lab-like picture book descriptions). The amount of dialectal exposure a child receives thus seems to depend on the region (more rural or urban), but also on the parental attitude toward their usage of dialect (cf. personal communication with families in our lab in Konstanz). Exclusively taking into account a child's residence is therefore an insufficient proxy for the amount of dialectal exposure it receives. Complementary perceptual judgments of dialectal input on the other hand seem a valid tool for the classification of a child's exposure to dialectal input (in addition to the Standard): Researchers have most often used rating scales with four categories (van Bezooijen and van Hout, 1985; Stölten and Engstrand, 2003; Floccia et al., 2009), but there are also studies that employed seven categories (Grondelaers et al., 2015), magnitude estimation (Brennan et al., 1975), or handgrip force (Brennan et al., 1975). For the purpose of this paper, children were divided into groups of dialectal vs. non-dialectal children according to the perceived dialectal strength of parental productions, which, as we will show, are correlated with the proportion of dialectal word forms that children are exposed to (see section "Participants"). The question emerging from the different amount of exposure to dialectal forms is

TABLE 1 | Example of differences between Standard German, Alemannic, and Middle Bavarian.

| | Standard German | Alemannic | Bavarian | English translation |
|---|----------------------------|-----------|----------|---------------------|
| 1 | [tʰ] [tʰɪʃ] (Tisch) | [dɪʃ] | [di:ʃ] | table |
| 2 | [s] [ʔo:pst] (Obst) | [ʔɔbʃt] | [ʔo:pst] | fruits |
| 3 | [l] [valt] (Wald) | [valt] | [vɔɪ̯t] | forest |
| 4 | [z] [diˈzʊnə] (die Sonne) | [ˈtsʊnə] | [tsɔn] | sun |
| 5 | [ə] [gəˈlaʊfən] (gelaufen) | [ˈglaʊfə] | [ˈglafə] | ran |
| 6 | [u:] [fu:s] (Fuß) | [fuəs] | [fuas] | foot |
| 7 | [y:] [ˈfʏ:lə] (Stühle) | [ˈftiəl] | [ftʊɪ̯] | chairs |

whether long-term exposure to word forms in more than one variety of a language affects children's word recognition abilities. The present study is designed to fill this gap. We will now turn to previous findings on word form recognition, particularly focusing on studies that test infants with exposure to more than one variety.

Word Form Recognition in Light of Dialectal Exposure

On their way toward learning words and building a vocabulary, one of the tasks children need to master is to acquire and refine word form representations (Westermann and Mani, 2018, for an overview). Children's ability to recognize word forms is commonly tested in the familiar word paradigm in which they are presented with two types of words: (familiar) words vs. nonce-words/rare words. As mentioned above, successful recognition of words typically surfaces in a preference for words over less familiar or nonce words (Carbajal et al., 2021). Children from a large number of different languages, including British and American English, Dutch, French, Spanish, Italian, and Japanese have been shown to recognize words from the end of the first year of life onward (26 experiments in Carbajal et al., 2021 tested children between 10 and 12 months of age), hence having started to develop lexical representations. The familiar word effect is influenced, among others, by children's age (stronger familiarity preference with increasing age), native language (stronger familiarity preference in Romance compared to Germanic languages), and their degree of word familiarity within real word lists (stronger familiarity preference when more familiar words were used). Under the assumption that the familiar word effect extends to German (a language not yet tested in this paradigm) and remote testing (previous studies were conducted in the lab), we predict a replication of the familiar word effect for non-dialectal German children aged 12–18-months (hypothesis H1). These children grow up with Standard German only which is why the presented Standard German word forms are assumed to be highly familiar to them. We explicitly included children older than 1 year of age (and hence older than children in most of the previous studies, cf. Carbajal et al., 2021) due to reasons of comparison between dialectal and non-dialectal children (for whom successful recognition of Standard forms might be observed later, see below). Another reason for testing 12–18-month-olds was that testing conditions outside the lab are different from typical laboratory settings, with less experimental control potentially leading to a reduction of the effect. This older age range might hence result in a more robust recognition effect.

In spoken communication, word forms are essentially variable, and children have to learn to recognize them in different (more or less variable) contexts (cf. White, 2018 for an overview). Indeed, it has been shown that children find it hard to recognize words when they are spoken with an *unfamiliar* accent. (Monolingual) children succeed in this task only by the end of the second year of life (Best et al., 2009; van Heugten and Johnson, 2014; van Heugten et al., 2018), suggesting very rigid early lexical representations that do not allow for deviations

from the form children are familiar with [but see Schmale et al. (2010) showing successful word recognition from around the first year of life in a different paradigm]. The situation, i.e., the mental representation of words, is probably different for children who grow up with two varieties of one language at a time. So far, however, we know very little about how long-term exposure to two varieties of a language affects the ability to recognize words, and the nature of early representations in these “bi-varietal” or dialectal children. For instance, children growing up in rural areas of Southern Germany are exposed to both Standard and dialectal forms (see section “Dialectal Variation in Germany and the Coding of Dialectal Input”), and regularly encounter both [fu:s] (Standard for “foot”) and [fūəs] (Alemannic for “foot”) in their input. Conceivably, the exposure to two varieties at a time affects how words are represented (see below).

Taken together, bi-varietal or dialectal upbringing leads to more variability in the input, but at the same time, also leaves the child with less input in either of the two varieties. Models of infant word recognition would generally predict that increased variability in the input is beneficial for the refinement of word forms and therefore facilitates the recognition of novel tokens (Johnson, 2016; see van Heugten and Johnson, 2017 for discussion). In this regard, speaker variability has indeed proven to be beneficial in word recognition and word learning (Singh, 2008; Rost and McMurray, 2009; Höhle et al., 2020). Little is known, however, about the effect of dialectal/varietal variability on word form recognition. There is one study by van Heugten and Johnson (2017) that investigated whether exposure to multiple accents affects the recognition of word forms. Specifically, the authors compared looking times to word lists containing familiar English words and nonce-word lists in children with low variability in the input (mainly exposed to Canadian English input, i.e., only one variety of English) vs. with high variability in the input (with around one third of exposure to Canadian English and two thirds of exposure to a different type of non-Canadian English, either another native English variety or foreign-accented English). Their results showed that while 12.5-month-old children from the low variability group successfully recognized Canadian English words, the high variability group only succeeded in this task at the age of 18 months. These findings hence suggest that exposure to multiple accents might in fact delay the familiar word effect rather than leading to beneficial processing. Based on these findings, we tentatively assume that the familiar word effect may surface later in our dialectal group as compared to their non-dialectal peers (hypothesis 2, H2). It needs to be mentioned though that the group of children tested in van Heugten and Johnson (2017) is more heterogenous compared to our group [all of our children are exposed to a native, Southern German dialect while children in van Heugten and Johnson (2017) are exposed to different types of native and non-native varieties], which might reduce direct comparability of the two studies.

While studies employing the familiar word paradigm may trace the development of word recognition abilities in different groups of children (e.g., mono- vs. bi-varietal children), they

cannot directly answer questions on the nature of lexical representations. This, however, is particularly relevant for children who grow up with two varieties. Bi-varietal or dialectal children may initially store (a) the form of one variety only (and thus only recognize the word forms of one variety, cf. Floccia et al., 2012), (b) the forms of both varieties (thus recognizing word forms of both varieties, cf. van der Feest and Johnson, 2016), or (c) develop underspecified representations (thus also accepting word forms with unattested phonological alternations, cf. Durrant et al., 2015). To answer such specific questions on the nature of lexical representations within the framework of the familiar word paradigm, one would need lists of words and corresponding nonce-words that are segmentally similar (e.g., all starting with the same consonant or having the same vowel) so that the reaction to a specific deviance can be tested. In this paper, we take a first step in this direction and use sets of words that share the same stressed vowel (both in the word and nonce-word tokens). For this purpose, we created two different word lists, one consisting of segmentally similar words that all contain the vowel [u:] as stressed vowel (u-only condition), and another consisting of segmentally varied words that contain mixed vowels in stressed position (u-varied condition). If the familiar word effect is comparable for mixed and segmentally similar word lists, future research could test the above-mentioned types of representations within this paradigm. For now, a successful recognition of Standard word forms in our dialectal group (which is expected to emerge later than in the monolingual group, cf. H2), would sustain the possibility of underspecified representations or double storage of forms in both varieties in dialectal children.

The remainder of this paper is structured as follows: In section “General Information on the App”, we first introduce the App that was developed and used to test a wide range of children in their home environments, and then describe the manual and semi-automatic coding that was used to analyze looking behavior. In section “Experiment 1: Word Form Recognition in 12–18-Month-Old Children,” we test the familiar word effect with German children between 12 and 18 months of age (Experiment 1). We manipulated dialectal input (between-subjects) as well as the nature of the materials (between-subjects) using different word lists (u-varied vs. u-only). Section “Experiment 2: Non-dialectal 18–24-Month-Old Children” tests non-dialectal children between 18 and 24 months of age (Experiment 2). In section “General Discussion,” we discuss the looking behavior and possible interpretations for word representations in dialectal and non-dialectal children. The same section concludes with discussing the applicability and limitations of the App used to study the familiar word effect with a more (linguistically and demographically) diverse population.

GENERAL INFORMATION ON THE APP

The App was developed by the fourth author (CZ). It is freely available in the App store¹.

¹<https://apps.apple.com/de/app/bsl-wortformerkennung/id1508534681>

Introduction of the App

In brief, a video introduces caregivers (mostly the mother) to the general procedure of the App, before they give consent and fill in a background questionnaire. In particular, we asked about the language(s) and dialect(s) the child is exposed to, and about the highest education of both caregivers as a proxy for socioeconomic status (Hoff et al., 2002; Bornstein et al., 2003; Ensminger and Forthergill, 2003; and references therein; Noble et al., 2005; Sirin, 2005). Furthermore, we asked whether the child is typically developed and whether there are any impairments in vision or hearing. The first phase of the experiment is a short production phase in which a caregiver describes a colorful picture to the child. The picture displays different people and animals, see **Figure 2A**. Parental speech input is used to judge the amount of dialectal variation a child receives (see section “Participants” for more details). The word recognition experiment itself starts with a calibration phase (an animated duck which appears in four corners of the screen) in order to establish reference points for manual coding (whether or not the child looks on the iPad or beyond its borders), see **Figure 2B**. The calibration phase is followed by the experimental trials of the word recognition experiment (see section “Procedure” for details). After the word recognition experiment, the caregiver is asked whether the other caregiver would like to describe the picture to the child again, whether there was any distraction during the word recognition experiment, and whether they would like to take part in a raffle. The data are then encrypted and securely transferred onto a password-protected university server for subsequent analysis.

Manual and Semi-Automatic Analysis of Looking Behavior

All eight experimental trials of each child were screened by one author (JK) to check the number of trials that a child completed. We included children who completed a minimum of six trials. To train the classifier for automatic coding, we selected two of the eight trial videos that differed most strongly in the child's orientation and/or movement. Looking behavior in those trial videos was coded manually frame-by-frame in ELAN (Brugman and Russel, 2004; ELAN, 2020), an annotation tool for video (and audio) recordings, as “look”, “no-look”, or “undecided”. All annotators were trained and received individual feedback on a set of videos that had previously been coded by two of the authors, both experienced in video annotation with ELAN (NC and KZ-R). Annotators received individual feedback on their annotations and training was completed once (a) the annotators felt confident in the coding process and (b) their annotations repeatedly did not differ more than \pm one frame from the boundaries in the annotation by the two authors that was used for baseline comparison, respectively. This was the case after no more than ten videos.

In total, four coders annotated the children's looking behavior. To determine the interrater agreement, we analyzed the coding of a coder pair frame-by-frame for a total of 24 videos (2 trials with a duration of 15 s from 12 children). The average agreement was 88.9%, Cohen's kappa 0.77, which is substantial (Landis and Koch, 1977), cf. **Table 2**.

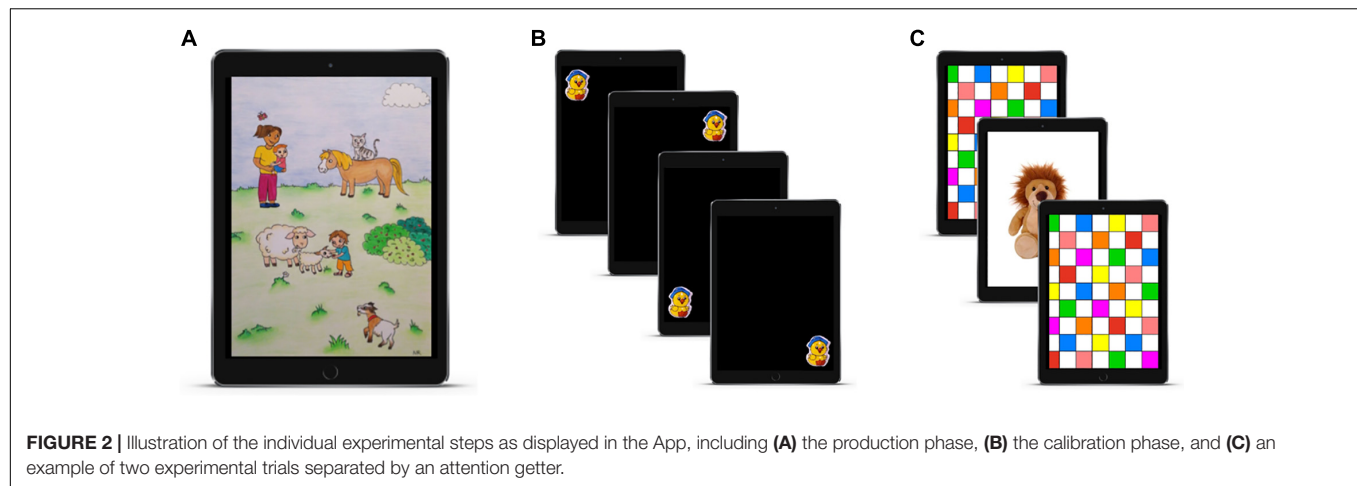


FIGURE 2 | Illustration of the individual experimental steps as displayed in the App, including (A) the production phase, (B) the calibration phase, and (C) an example of two experimental trials separated by an attention getter.

Using the manually annotated trials, a semi-automatic annotator was trained to process the remaining videos. The semi-automatic annotation is a two-step process: In the first step, parameters are extracted from the videos, using face and landmark detection software from 4dface². In the second step, each frame is classified using a Long-term Recurrent Convolutional Network (cf. Donahue et al., 2015 for successful application in human action classification).

For parameter extraction, all faces which are visible in the video are detected using the deep learning-based face detector from 4dface and the child's face is selected either automatically based on position (lower on screen, more centered) or manually in ambiguous situations (e.g., presence of a sibling). Then, the face is tracked over the duration of the video and the facial landmarks are localized. Seven parameters are calculated for each frame: the face's orientation (pitch, yaw, and roll), the relative x and y position to the center of the image, the distance between the outer corners of the eyes and the distance between the chin and the center point between the eyes. Additionally, a cropped image of each eye is saved for each frame. For frames in which the face is not detected, placeholder values are saved instead.

For the classification step, we constructed and trained a Long-Term Recurrent Convolutional Network (Donahue et al., 2015), which combines the previously calculated numeric parameters with the eye images and returns a label for each frame. The LSTM is capable of incorporating temporal context to the classification, so the input to classify one frame is not only based on the parameters of the frame itself, but also on the seven frames before and after. This improves classification of frames in which the

eyes are not visible because of occlusion, or frames in which the child is blinking. A simplified representation of the model can be seen in Figure 3. In the network, the eye images are fed through a Convolutional Neural Network (ConvNet), which learns to interpret the images and represents them in a dense layer. The dense layer is concatenated with the other numeric parameters, which are then fed into the LSTM. The LSTM assigns the label "look" or "no look" to classify the middle frame; the "undecided" category is ignored during training because it does not contain enough data points. To deal with outliers, the resulting series of frame classifications is smoothed. This corrects some erroneous classifications, which are mostly single frames that are classified with a different label than the ones surrounding it.

For training the LSTM, the manually annotated videos are split in a training (75%) and validation set (25%), making sure that at least one video of each child occurs in the training set. Ten videos are held back as a test set. The training set is augmented with a mirrored version of each video. Using drop-out and kernel regularization on the dense layers is essential to prevent overfitting. The final model achieves an average agreement between the manual annotations and the semi-automatic annotations of 97% on the training set, 94% on the validation set and 93% on the test set. The average Cohen's kappa between the manual annotations and the semi-automatic annotations is 0.83 on the test set.

EXPERIMENT 1: WORD FORM RECOGNITION IN 12–18-MONTH-OLD CHILDREN

Experiment 1 is a replication of the familiar word paradigm to test word form recognition of German non-dialectal and dialectal 12–18-month-olds, using the App.

Methods Participants

The grouping of children into a non-dialectal and a dialectal group was based on the parental recordings (gathered in the

²<https://4dface.io/>

TABLE 2 | Pairwise interrater agreement for the pairs of coders for look/no-look.

| Annotator pair | Agreement |
|-----------------------|-----------|
| Annotator1-Annotator2 | 92.3% |
| Annotator2-Annotator3 | 86.6% |
| Annotator2-Annotator4 | 81.3% |
| Annotator3-Annotator4 | 95.4% |

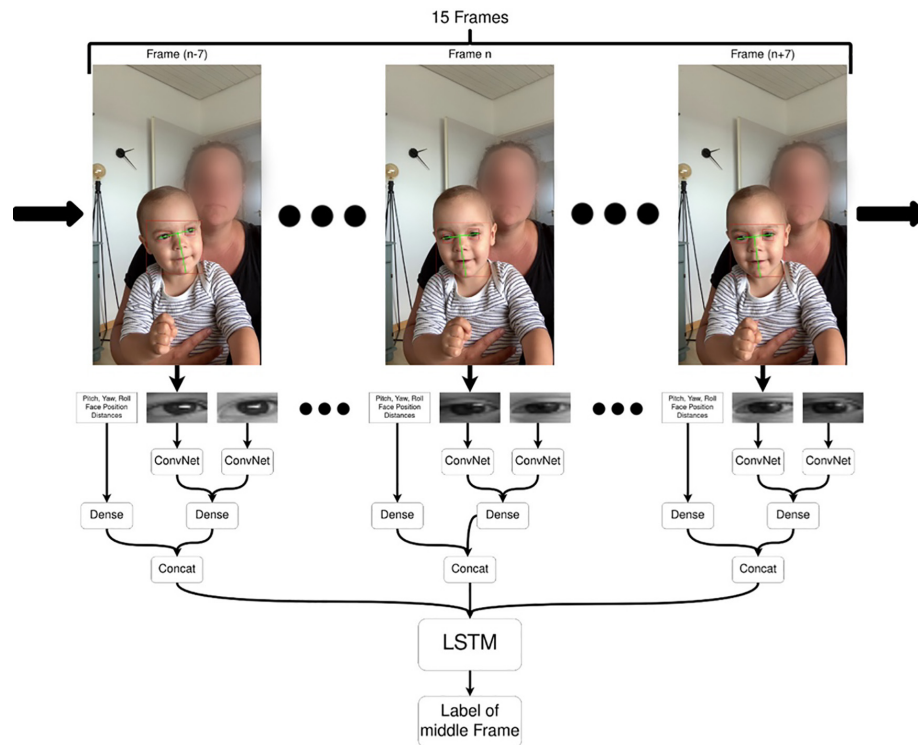


FIGURE 3 | Overview of the automatic coding procedure.

production phase, see **Figure 2A**). To this end, the input data were converted to the wav media format and each rated by four student assistants with respect to dialectal strength. The student assistants were all students of linguistics (at least in their 2nd year). There were two fixed sets of four student assistants. The student assistants were selected based on their place of origin in Germany to reduce effects of familiarity with some of the dialects. The coders rated the dialectal strength on a 4-point Likert scale. Perfectly Standard German stimuli were coded as 1, stimuli with a few slight dialectal features as 2, with more dialectal features as 3, and 4 was used for highly dialectal productions, see **Supplementary Table 1** for examples of dialectal productions. We calculated Cronbach's α as a measure of reliability (Cronbach, 1951); α was 0.94 for the first set of annotators and 0.92 for the second set. We used the mode (most frequent rating among the four coders) to group children into two dialect groups. We used a mode of 2 as the cutoff-value. Non-dialectal children had a mode of 1 or 2, dialectal children had a mode of 3 or 4. There were four critical ties (mode 2 vs. 3), which were resolved by soliciting an additional rating of a randomly selected rater from the other group.

To support the classification into dialectal and non-dialectal children, a random selection of picture descriptions (four from the non-dialectal and four from the dialectal group) were coded phonetically, following the procedure described in Zahner et al. (2021). More specifically, each word form was coded according to its segmental deviation from the expected Standard German form due to dialectal variation, e.g., [nø:t] for [niçt]

nicht "not," or due to general reduction processes that occur in connected speech, e.g., [niç] for [niçt] (Kohler, 1990). The proportion of dialectal word forms was 9% (SD = 2%) for the caregiver of the non-dialectal group and 34% (SD = 8%) for the dialectal group. The average dialectal strength (averaged over four rates) was highly correlated with the proportion of word forms that contain dialectal deviances (Spearman's $\rho = 0.88$), see **Supplementary Figure 1**. This emphasizes the relation between the perception of dialectal strength and the frequency of occurrence of dialectal variants.

Forty-four children were included in the analyses of Experiment 1, 25 non-dialectal children (13 in the u-varied condition, 12 in the u-only condition) and 19 dialectal children (11 in the u-varied condition, 8 in the u-only condition). They were matched for gender and age (see **Table 3** for details). Eleven of the children came from a Swabian area (Zip-Code 72, north of Konstanz), five from an Alemannic area (Zip-Code 78 and 79, Bodensee area) and three from Bavaria, south of Munich (Zip-Codes 82 and 83). Seven more children were tested, but not analyzed (4 non-dialectal and 3 dialectal) because the child did not pay attention (3 times), was not in the frame (2 times), was reported to be ill or impaired (1), or there was loud background noise (1). This resulted in a dropout rate of 13.7%.

Materials

Selection of Words and Nonce-Words

Eighteen frequent German words were selected, twelve words with the vowel /u/ and six words with mixed vowels [other long

TABLE 3 | Overview of participant metadata.

| | Non-dialectal children | | Dialectal children | |
|--|------------------------|------------|--------------------|------------|
| | u-varied | u-only | u-varied | u-only |
| Gender (number of female/male) | 6/7 | 7/5 | 5/6 | 3/5 |
| Mean age (SD) in months | 14.5 (1.5) | 15.6 (1.9) | 14.5 (2.1) | 14.4 (1.9) |
| Mean dialect score | 1.4 (0.4) | 1.6 (0.4) | 3.3 (0.5) | 3.2 (0.3) |
| Highest education of caregiver 1 (university degree/vocational training/A-levels/O-levels) | 9/2/2/0 | 9/2/1/0 | 4/5/1/1 | 3/2/3/0 |

vowels (/a:/ and /i:/), other short vowels (/a/ and /œ/), and the diphthong [äʊ], see first two columns of **Table 4**.

All selected words had at least one consonant in the onset position of the first syllable. Nine of the words were monosyllabic (e.g., *Stuhl* [ʃtu:l], “chair”) and nine were disyllabic with a trochaic stress pattern (e.g., *Katze* [kʰat͡sə], “cat”). They were all expected to be known by German 24-month-olds. They had a log-lemma frequency count higher than 0.9 from dlexDB (Heister et al., 2011) with an average of 1.85, see third column of **Table 4**. Furthermore, they are produced by at least a quarter of German 24-month-old children (average 69%), as indicated in Wordbank (Frank et al., 2017), an open database of children’s vocabulary growth. The Wordbank (WB) database³ collects MacArthur-Bates Communicative Development Inventory (MB-CDI) data in many different languages, i.e., information from parent-report questionnaires on children’s vocabulary growth (Szagun et al., 2009). The German data set consists of 1,181 children aged between 18 and 30 months. The retrieved WB data (fourth column of **Table 4**) shows the proportion of German children producing an item at a specific age. The words in the u-only and u-varied lists did not differ from each other in terms of frequency and word bank data (see **Supplementary Table 2**).

Nonce-words were constructed in the following way: For disyllabic words, we exchanged the onset consonants of the syllables of each word with those of another word (e.g., [ʰu:s.tʰən] → [bʁu:.xən]). For monosyllabic words, we exchanged onset and coda consonants between word pairs ([bɑʊm] → [hɑʊl]). Each word formed a pair with its newly constructed nonce-word, see **Table 4**, last two columns. This procedure ensured that the pairs of words and nonce-words were comparably complex regarding consonant clusters and syllable structure. Consonants were exchanged because they strongly affect word recognition (Pollock and Nazzi, 2015). One important criterion for the nonce-word generation was that phonotactic probabilities were matched for words and nonce-words (Vitevitch and Luce, 2004, p. 481). Similar to the web-based Phonotactic Probability Calculator for English⁴ [used, e.g., in van Heugten et al. (2018)], we extracted positional segment frequencies and position-specific biphone frequencies for each (nonce-) word from the wordform dictionary of the CELEX lexical database (Baayen et al., 1995) using a self-programmed

TABLE 4 | List of words and their IPA-transcription (first two columns) and the generated nonce-words (last two columns).

| Word | IPA | dlexDB | WB 18 m/24 m | Nonce- word | IPA (Standard) |
|-----------------------------------|-------------|--------|-----------------|----------------|-------------------|
| <i>Kuchen</i> “cake” | [kʰu:.xən] | 0.98 | 0.16/0.66 | <i>Buten</i> | [bʉ:.tʰən] |
| <i>Fuß</i> “foot” | [fu:s] | 2.07 | 0.27/0.75 | <i>Stuch</i> | [ʃtu:x] |
| <i>Kuh</i> “cow” | [kʰu:] | 1.18 | 0.22/0.8 | <i>Fuh</i> | [fu:] |
| <i>Stuhl</i> “chair” | [ʃtu:l] | 1.70 | 0.19/0.71 | <i>Guhm</i> | [gu:m] |
| <i>Schuh</i> “shoe” | [ʃu:] | 1.45 | 0.45/0.78 | <i>Kud</i> | [kʰu:tʰ] |
| <i>Buch</i> “book” | [bu:x] | 2.34 | 0.39/0.86 | <i>Zust</i> | [t͡su:st] |
| <i>Blume</i> “flower” | [ʰblu:.mə] | 1.65 | 0.28/0.72 | <i>Bluche</i> | [ʰblu:.xə] |
| <i>gut</i> “good” | [gu:tʰ] | 3.05 | 0.12/0.47 | <i>Suh</i> | [zu:] |
| <i>Bruder</i> “brother” | [bʁu:.dɐ] | 2.02 | 0.02/0.25 | <i>Schuser</i> | [ʃu:.zɐ] |
| <i>Husten/husten</i> “(to) cough” | [ʰu:s.tʰən] | 1.28 | —/— | <i>Bruchen</i> | [bʁu:.xən] |
| <i>zu</i> “to” | [t͡su:] | 3.96 | 0.28/0.57 | <i>Hu</i> | [hu:] |
| <i>suchen</i> “to search” | [ʰzu:.xən] | 2.37 | —/— | <i>Kulen</i> | [kʰu:.lən] |
| <i>Hase</i> “hare” | [ʰha:.zə] | 0.91 | 0.27/0.73 | <i>Kafe</i> | [kʰa:.fə] |
| <i>Baum</i> “tree” | [baʊm] | 1.87 | 0.26/0.73 | <i>Haul</i> | [haʊl] |
| <i>spielen</i> “to play” | [ʃpi:.lən] | 2.32 | 0.13/0.58 | <i>Biesen</i> | [bi:.zən] |
| <i>Katze</i> “cat” | [kʰat͡sə] | 1.29 | 0.26/0.76 | <i>Lamme</i> | [lamə] |
| <i>Ball</i> “ball” | [bal] | 1.20 | 0.76/0.95 | <i>Spall</i> | [ʃpal] |
| <i>Löffel</i> “spoon” | [lœfəl] | 1.05 | 0.19/0.71 | <i>Bötzel</i> | [bœt͡səl] |

The first 12 pairs are used in the u-only condition, the first six and last six pairs in the u-varied condition. Segmental changes are highlighted in bold face. The middle columns give information on lexical frequency and production frequency at 18 months and at 24 months of age. Numbers in *italics* were only available for the plural form.

Python script. **Table 5** shows that the mean phonotactic probabilities are matched at the segment and biphone level⁵.

Acoustic analyses

A 26-year-old female native speaker of Standard German from the Southwest of Germany (Baden-Wuerttemberg) recorded the thirty-six experimental items in isolation (18 words and 18 nonce-words). She was instructed to produce them as if naming them for a small child, resulting in (rising)-falling intonation contours. Words and nonce-words were closely matched according to a number of acoustic parameters, i.e., duration of the target word, duration of the stressed syllable, mean f0 in stressed syllable, and f0 excursion of the accentual fall in the target (H* L-%), see **Supplementary Table 3**.

⁵One reviewer pointed out that some of the nonce-words may be perceived as a mispronunciation of an existing word. In the literature, children from 11 months onward do not recognize words in which individual sounds (or features) are changed, especially for consonants (e.g., Swingley, 2005; Pollock and Nazzi, 2015). To test whether the nonce-words may be understood as existing German words, we played all nonce-words to a group of eight student assistants. They were given 2 s to name a word that spontaneously came to their mind or to respond with “X” if no word occurred to them. Only four nonce-words led to associations with an existing word, whereby more than half of the student assistants, respectively, mentioned the same word. Overall, we find this proportion (or amount of deviance between words and nonce-words) to be comparable to other studies using the familiar word paradigm (Vihman et al., 2004; Swingley, 2005; van Heugten and Johnson, 2014; Vihman and Majorano, 2017).

³<http://wordbank.stanford.edu/>

⁴<https://calculator.ku.edu/phonotactic/English/words>

TABLE 5 | Mean phonotactic probabilities (and standard deviations) of words and nonce-words.

| | u-only | | u-varied | |
|----------|-------------|-------------|-------------|-------------|
| | Words | Nonce-words | Words | Nonce-words |
| Segments | 1.22 (0.13) | 1.22 (0.13) | 1.20 (0.09) | 1.22 (0.10) |
| Biphones | 1.02 (0.03) | 1.03 (0.03) | 1.02 (0.02) | 1.02 (0.02) |

The words and nonce-words were also matched for speaker affect. To this end, all twelve words and nonce-words from the u-varied list were presented together with twelve less emphatic recordings of the same word and the same speaker (not used as experimental stimuli but recorded for the purpose of the rating task). Ten listeners rated these tokens (which occurred in both the u-varied and u-only condition) on a scale from 1 (= not enthusiastic at all) to 5 (= very enthusiastic). The words received an average rating of 3.92 (SD = 0.71), the nonce-words an average rating of 3.89 (SD = 0.76), corroborating that the words and nonce-words do not differ in perceived speaker affect.

Experimental Lists

The recordings of the experimental tokens were concatenated into four word lists and four nonce-word lists for both the u-varied and the u-only condition. The lists only differed in the order of tokens. Following van Heugten et al. (2018), the order of (nonce-)words varied within each list. Every token appeared only once. Moreover, no more than two monosyllabic or bisyllabic (nonce-)words occurred immediately adjacent. Across lists, each token appeared in early, mid and late positions of the list. Two of the four lists of each vowel-type were mirror lists of each other. Each list hence contained all twelve tokens (word or nonce-word tokens, respectively). The tokens were equated for loudness (65 dB) and concatenated with silent inter-stimulus intervals (ISIs) of approx. 750 ms, see van Heugten et al. (2018). To ensure an equal duration of all lists, the ISIs were adapted to reach list durations of 15 s (ISI was the same for each list and ranged from 742 to 756 ms). Each child received all four word lists and all four nonce-word lists of one vowel-type (vowel type was manipulated between-subjects). We constructed six different experimental randomizations of the above lists (i.e., of eight trials each), such that word lists and nonce-word lists did not appear more than two times in a row. Moreover, in all randomizations of the lists, word and nonce-word lists were balanced for experimental half, i.e., two of the four word lists and two of the four nonce-word lists occurred in one experimental half. Three of the six randomizations started with a word list, while three started with a nonce-word list. Randomizations were the same for u-varied and u-only conditions.

Procedure

The remote familiar word paradigm consisted of eight trials in total (four word lists and four nonce-word trials, 15 s each). Each trial started with a colorful attention getter (taken from Frota et al., 2014), which was presented for 1 s (cf. Figure 2C). Then one of the word or nonce-word lists was played, accompanied by the visual presentation of a colored

checkerboard. The sound played for the total duration of the list (and was hence not child-controlled). For each experiment version (u-only and u-varied), the six different randomizations of trials described above were distributed across participants. For the analysis of looking times, two of the trials of each child were coded manually on a frame-by-frame basis for looks in ELAN. The analysis was based on the automatic coding (cf. see section “Manual and Semi-Automatic Analysis of Looking Behavior”).

Results

The looking times were slightly left-skewed, which is why we transformed them using a square-root transformation, see Eq. 1. The negative sign ensured that longer transformed looking times correspond to longer raw looking times.

$$\text{transformed_lookingtime} = -\sqrt{16,500 - \text{lookingtime}} \quad (1)$$

The transformed looking times were analyzed in a linear mixed-effects regression model with *group* (non-dialectal vs. dialectal, treatment-coded), *word-type* (word vs. nonce-word, treatment-coded), *vowel-type* (u-only vs. u-varied, sum-coded) and the control predictors *block* (first vs. second block of trials, sum-coded) and scaled *age in months*. Block was included to test whether looking time differences are already present at the onset or develop over the course of the experiment, due to exposure to the stimuli (cf. analysis in Poltrock and Nazzi, 2015). Sum coding of predictors allowed us to focus on the main factors of interest, *group* and *word-type* in the summary()-tables. *Participants* were added as random effect (Baayen et al., 2008). Adding experimental list as random effect did not lead to model convergence. If the model converged, *block* and *vowel-type* were added as random slopes for participants. Due to convergence issues, only *block* was kept as random slope. The final model showed a significant four-way-interaction between *word-type*, *group*, *block* and *age* [$F_{(1, 255)} = 10.0, p < 0.005$], see Figure 4 for marginal effects of the model. There was no effect of *vowel-type* and no interaction between *vowel-type* and any of the other factors ($p > 0.1$). A Bayes factor analysis (Morey and Rouder, 2018) showed that the model without *vowel-type* as predictor was more than 10,000 times more likely than the model with *vowel-type*. We investigated the four-way-interaction more closely by fitting separate models for the non-dialectal and dialectal groups.

Figure 5 shows the raw looking time differences per child (panel A for non-dialectal, panel B for dialectal children) and the raw looking times per trial (panel C for non-dialectal, panel D for dialectal children). The non-dialectal group showed a main effect of *word-type* [$F_{(1, 146)} = 7.4, p < 0.01$], with longer looking times to word lists than nonce-word lists ($\beta = 7.9, SE = 2.9$) and of *block* [$F_{(1, 23)} = 19.9, p < 0.0005$], with longer looking times in block 1 than block 2. The effect size (Hedge's *g*) for the effect of *word-type* was 0.43, 95% CI [0.11;0.76] (small effect). Furthermore, there were significant interactions between *word-type* and *age* [$F_{(1, 146)} = 10.2, p < 0.005$], and between *word-type*, *age* and *block* [$F_{(1, 146)} = 7.2, p < 0.01$], see left panel of Figure 4. The familiarity preference decreased with increasing age and was more pronounced in block 2, in particular for

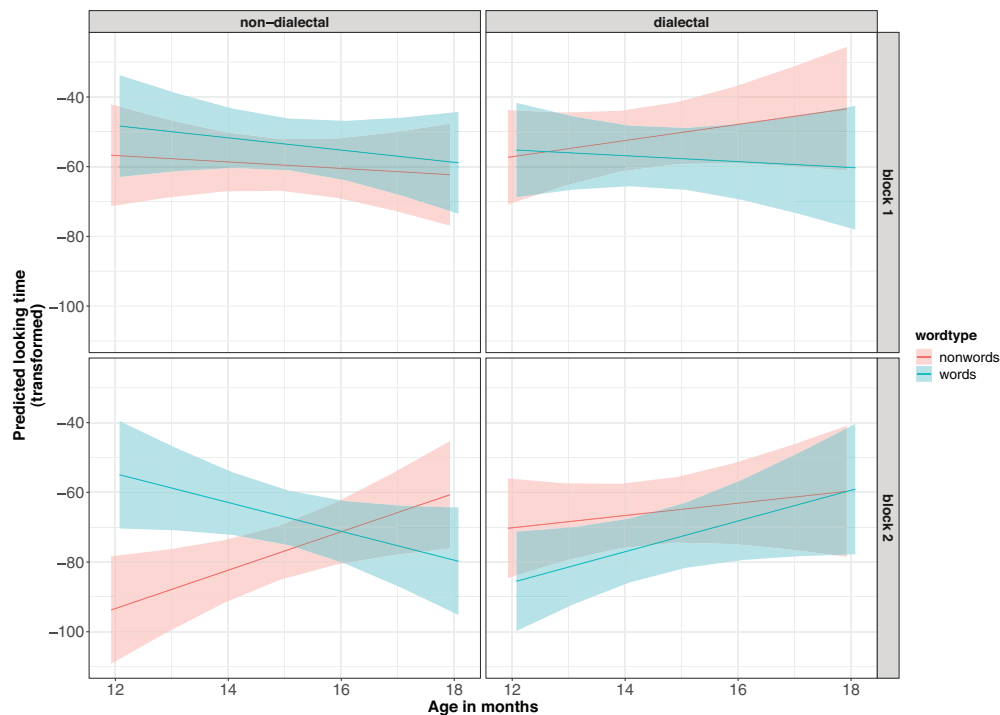


FIGURE 4 | Predicted effects of the model in Experiment 1. The y-axis shows the transformed looking times – $\sqrt{16,500 - \text{lookingtime}}$. Higher values indicate longer looking times.

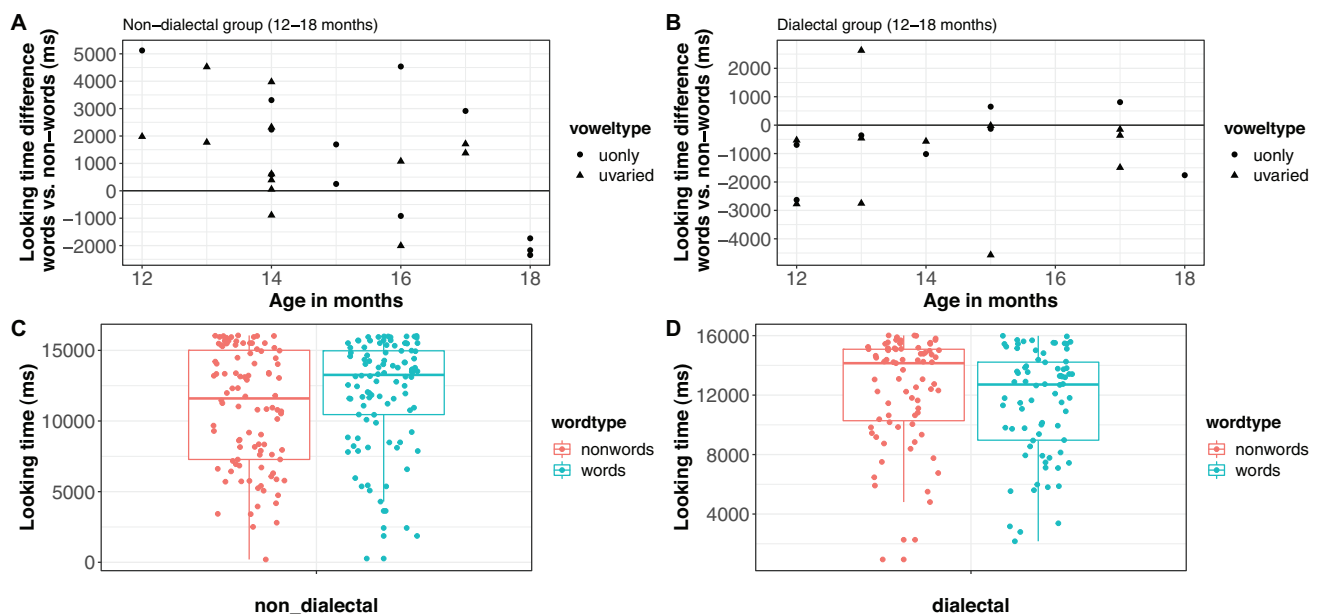


FIGURE 5 | Looking time difference between words and non-words in Experiment 1 [(A,B): each dot represents the average of one child] and looking times by word type [(C,D): each dot represents one trial]. Left panels: non-dialectal children, right panels: dialectal children.

the younger children. Separate analyses by block showed longer looking times to words than non-words in block 1 [$F_{(1, 72)} = 3$, $p = 0.08$] and an interaction between *word-type* and *age* in block 2 [$F_{(1, 73)} = 20.2$, $p < 0.0001$], see **Figure 4**, left panel.

The dialectal group showed significant main effects of *word-type* [$F_{(1,127)} = 4.7$, $p < 0.05$], with longer looking times to non-words than to words ($\beta = 7.1$, $SE = 3.3$) and of *block* [$F_{(1,127)} = 25.1$, $p < 0.001$], with shorter looking times in

block 2. Furthermore, there was a three-way-interaction between block, *vowel-type* and *age* [$F_{(1,127)} = 6.3, p < 0.05$]. The effect size (Hedge's *g*) for *word-type* was 0.34, 95% CI [$-0.64; -0.04$] (small effect).

Discussion

Experiment 1 showed looking time differences between word and nonce-word lists for both non-dialectal and dialectal children. However, the preference went in opposite directions: Non-dialectal children preferred words over nonce-words, while dialectal children preferred nonce-words over words. With regard to the non-dialectal children, who primarily receive Standard German input, we hence replicated the familiar word preference with German children in a home-setting using an App. We further showed that the familiar word preference was not affected by segmental variation of the stimuli (u-varied vs. u-only), but the familiarity preference was stronger in block 2, in particular for the younger children. The familiarity preference for words over nonce-words is in line with a number of studies that tested children in different languages in the lab (Swingley, 2005 for Dutch; Best et al., 2009 for American English; van Heugten and Johnson, 2014 for Canadian English; Poltrock and Nazzi, 2015 for French; Vihman and Majorano, 2017 for Italian, among others). At 18 months, the pattern seems to slowly develop into a novelty preference (Figure 5A). A similar decline in familiarity preference was also observed in Vihman et al. (2007). We will come back to this reversal of preferences in Experiment 2.

The dialectal children, who are exposed to more variability in word forms (both dialectal and Standard German word forms), showed a preference for the nonce-word lists, suggesting that they also recognized the Standard German word forms. The direction of the preference, however, is rare in the literature on this paradigm. To be more confident about the obtained effect in dialectal children, which was indeed unexpected, we challenged its stability by removing the child with a particularly large novelty preference ($> 4,000$ ms); this looking time difference was 2.3 SD beyond the mean and does hence not qualify as an outlier in a strict sense. Moreover, we reran the analysis with an adapted looking time measure, excluding looks that occurred after a sequence of “no look” frames longer than 2 s (to simulate a child-controlled paradigm one would have used in the lab). This affected about half of the trials ($N = 70, 46\%$), nevertheless the pattern of results did not change. In both cases, the novelty effect persisted, which corroborates its robustness.

Novelty preferences have been associated with more mature linguistic processing in the literature. DePaolis et al. (2016), for instance, showed that within one and the same age group (10-month-old children), successful recognition of familiar words can surface either as a familiarity preference or a novelty preference. Children who showed a preference (familiarity or novelty) in that study were lexically more advanced (as measured by standardized vocabulary assessments, CDI, MacArthur Communicative Developmental Inventory) than the children who did not show a preference (equal looking times to both stimuli types). The novelty preference in dialectal children in Experiment 1 may hence be tentatively interpreted as an effect

of more mature linguistic processing. To test this hypothesis, we looked at a sample of older non-dialectal children, which are, naturally, more mature than the 12–18-month-olds, and for whom one may expect a similar novelty preference (cf. Thiessen et al., 2005 for a similar rationale in word segmentation)⁶. For the familiar word paradigm, there are only very few studies with children older than 19 months (Carbajal et al., 2021), who are typically tested with non-native accents or with specific populations (Best et al., 2009; van Heugten and Johnson, 2014; Kalashnikova et al., 2016). van Heugten and Johnson (2014) report an interaction between word type and age and suggest that “over time, infants start preferring to listen to known over nonsense words” (p. 344). Kalashnikova et al. (2016) tested a control group of 26-month-old children with familiar accents and showed a reduction of the familiar word preference for 26-month-old children.

EXPERIMENT 2: NON-DIALECTAL 18–24-MONTH-OLD CHILDREN

Experiment 2 follows up on the different directions of preferences observed in Experiment 1 (familiarity preference in non-dialectal vs. novelty preference in dialectal children). If the novelty preference is indeed indicative of more mature processing and if the familiar word paradigm is still a valid method for 18–24-month-old children, we would expect a change in the direction of the preference toward a novelty effect in non-dialectal children as they grow older.

Methods

Participants

Twenty non-dialectal children between 18 and 24 months of age were included in the analysis (10 female and 10 male). Their mean age was 20.6 months ($SD = 1.6$ months). Ten children were tested with the u-varied lists, ten with the u-only lists. Their mean dialect score was 1.5 ($SD = 0.4$). The highest education of the first parent equaled a high school degree (for 13 children), vocational training (5 children), A-levels (1 child), and O-levels (1 child). Eight more children were tested, but not analyzed because the child did not complete the test (2 times), was not in the frame (2 times), was reported to be ill or impaired (3 times) or because of technical issues (1). This resulted in a dropout rate of 28.6%.

Materials and Procedure

The materials and the procedure were identical to Experiment 1.

Results

The raw looking time differences per child and looking times per trial are shown in Figure 6. The looking times were transformed and analyzed as in Experiment 1. The final model showed significant effects of *word-type* [$F_{(1, 137)} = 6.6, p < 0.05$], *age* [$F_{(1, 17)} = 8.4, p < 0.05$], *vowel-type* [$F_{(1, 17)} = 5.4, p < 0.05$], and *block* [$F_{(1, 137)} = 19.1, p < 0.001$]. Furthermore, there was

⁶There were not enough data points of dialectal 18–24-month-olds for comparison, unfortunately.

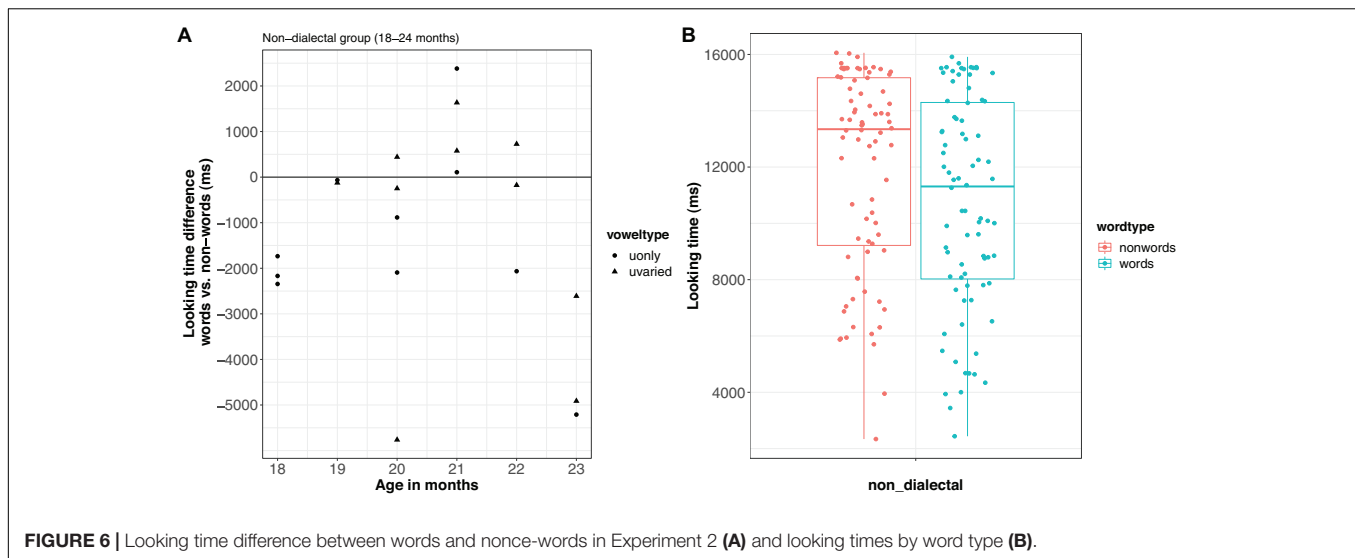


FIGURE 6 | Looking time difference between words and non-words in Experiment 2 (A) and looking times by word type (B).

an interaction between *vowel-type* and *block* [$F_{(1, 137)} = 4.9$, $p < 0.05$]. Importantly, children looked longer to the non-words lists than to the word lists ($\beta = 9.1$, $SE = 3.5$, Hedge's $g = 0.51$, 95% CI $[-0.96; -0.07]$, medium effect). Furthermore, children looked longer to the u-varied lists than to the u-only lists ($\beta = 21.1$, $SE = 6.7$), especially in block 1. Looking times were furthermore shorter for older children ($\beta = -8.3$, $SE = 2.9$) and in the second block ($\beta = -7.7$, $SE = 5.0$).

Figure 7 shows looking time differences for non-dialectal children across age in one figure. It demonstrates how the familiarity preference (overserved in Experiment 1) slowly develops into a novelty preference (Experiment 2) as a function of age.

Discussion

In Experiment 2, non-dialectal 18–24-month-olds preferred non-words over words, suggesting that non-dialectal children

indeed develop a novelty preference as they grow older, with longer looking times to non-words lists than to word lists. This tendency toward a novelty preference with increasing age has already been foreshadowed in non-dialectal children in Experiment 1: There, the familiarity effect was largest for the younger children in that group. It is not surprising that older children (18–24-month-olds) are developing a growing interest in non-words. At this age, the vocabulary develops very rapidly (e.g., Dapretto and Bjork, 2000) and an interest in novel words is the best way to further increase a child's lexicon. Importantly, the looking behavior of older non-dialectal children resembles the behavior of the 12–18-month-old dialectal children who grow up with dialect forms in addition to Standard German (see Experiment 1). The fact that older non-dialectal and younger dialectal children both show a novelty preference does not necessarily mean that the cause is of the same origin. One generalization we can still infer from our findings is linguistic maturation (caused by a more variable input due to Standard German and dialectal forms in the dialectal group of Experiment 1 or caused by increased age in Experiment 2). We will further discuss this interpretation in the “General Discussion” section.

GENERAL DISCUSSION

The present paper tested word form recognition in German children aged 1–2 years, growing up with Standard German (non-dialectal children, mostly recruited in urban areas) or with Standard German *and* an additional dialect (dialectal children, mostly recruited in rural areas). Data collection was made possible by an App that allowed parents to run the experiment at home, using an experiment-controlled version of the familiar word paradigm. As predicted by H1, non-dialectal German 12–18-month-old children showed the preference for familiar over non-words established in the literature (cf. Carbajal et al., 2021). The familiarity preference was stronger for younger children, in particular in the second half of the experiment. From 18 months of age onward, this familiarity preference

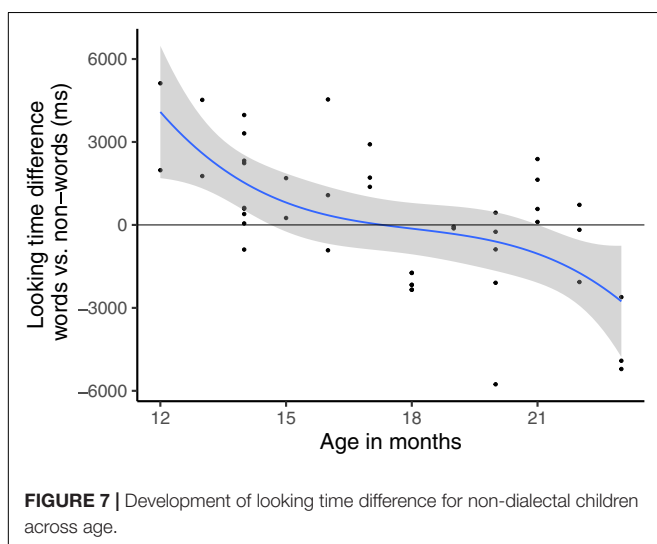


FIGURE 7 | Development of looking time difference for non-dialectal children across age.

slowly developed into a novelty preference, with longer looking times for nonce-word than word lists. The task might have become successively simpler which increases the likelihood of a novelty preference and an increased interest in novel items. These results extend earlier findings on the familiar word preference for a different language (German) and for an older age group (12–18 months). Our main focus, however, lies on the dialectal German children, who receive a lot of variability in their daily input. The group of dialectal 12–18-month-olds showed a novelty preference comparable to older non-dialectal children. Since novelty preferences are rare within this paradigm, this finding was particularly unexpected (recall that based on the literature we had assumed a later occurrence of a familiarity preference for the dialectal group as compared to the non-dialectal group, cf. H2). We therefore critically assured that the novelty preference is indeed statistically robust. Finally, we concluded that this preference pattern is most likely to be attributed to more advanced linguistic skills in dialectal children, due to experience with variability in word forms. The segmental variability of the word (and nonce-word) lists did not affect the familiarity or novelty preferences. This suggests that specific alternations can be tested in the familiar word paradigm, which allows us to investigate the nature of early word representations with this paradigm (see section “Direction of the Preference”).

In the following, we first reflect on the classification into dialectal vs. non-dialectal children (see section “Operationalization of Dialectal vs. Non-dialectal”). Then, we discuss the direction of the preference (see section “Direction of the Preference”) before turning to the nature of lexical representations in dialectal children (see section “Effects of Dialectal Exposure on Lexical Development”). We conclude with an evaluation of the remote testing procedure using the App (see section “Evaluation of the Remote Testing Using the App”).

Operationalization of Dialectal vs. Non-dialectal

In a first attempt to study the role of dialectal variability in children's word form recognition, we used a binary classification procedure. This binary classification of children into a dialectal vs. non-dialectal group was primarily based on a perceptual measure, i.e., the impression of dialectal strength by a group of four coders. The coders came from different areas of Germany in order to avoid that familiarity with a particular dialect skewed the ratings in any form. Our coding system of dialect strength proved to be fairly reliable, which is in line with other studies that also reported high interrater reliability for the perceptual coding of dialect strength, even among lay coders (Ryan, 1973: Kendall's $W = 0.71$; van Bezooijen and van Hout, 1985: $\kappa = 0.93$; Grondelaers et al., 2015: $\alpha = 0.6$). This suggests that native speakers of a language are able to reliably perceive and indicate the strength of dialectal usage. Furthermore, previous studies have established that subjective measures of dialectal strength and objective phonetic measures are highly correlated (e.g., euclidian distance of first and second vowel formants from a reference, cf. Grondelaers et al., 2015), which further corroborates the validity of such coding systems.

In our study, the ratings reflect the perceived prevalence of dialectal deviations from expected Standard forms (recall that perfectly Standard German stimuli were coded as 1, stimuli with a few slight dialectal features as 2, with more dialectal features as 3, and 4 was used for highly dialectal productions). Although it is unlikely that the coders tracked these frequencies in an accurate one-to-one manner, phonetic transcriptions of the realizations of the subset of the recordings showed a very high correlation (Spearman's $\rho > 0.88$) between the perception of dialectal strength and the number of word forms that deviated from Standard German forms. Taken together, our system seems to have reliably grouped children into non-dialectal vs. dialectal children. Nevertheless, more fine-grained approaches are conceivable.

Our binary grouping aimed at serving as a first approximation toward operationalizing the dialect-induced variability in word forms, which allowed us to investigate the development of lexical representations. However, given that the use of dialect ranges over a continuum (Schwarz, 2014), and that present-day dialectal forms approach Standard German word forms (e.g., Kleber, 2020), it may be more valid to include dialectal strength as a continuous measure in the analyses for future studies, cf. Levy et al. (2019) for the role of experience in the processing of accented speech (unfamiliar regional and foreign) in 9-year-old children; and Porretta et al. (2016) on the influence of foreign accentedness and experience in adult word recognition. In addition, it may be helpful to include more questions on the frequency of dialect use and a self-assessed rating of dialect strength, similar to questionnaires used for bilingual children (e.g., Levy et al., 2019; DeCat, 2020).

Direction of the Preference

The two groups tested in our study revealed effects of different directions. Almost all of the previous studies using the familiar word paradigm showed a familiarity preference [$N = 18$ out of 32, see lower half of Figure 2 in Carbajal et al. (2021)] if there was an effect, rather than a novelty preference. Most of the children in the familiar word paradigm were younger than 12 months of age (25 out of 32 studies in Carbajal et al., 2021). There is only one study that showed a novelty preference for linguistically more mature children (as measured by higher CD scores), cf. DePaolis et al. (2016). This study, along with previous reflections on the direction of effects (Hunter and Ames, 1988; Houston-Price and Nakai, 2004; Butler et al., 2011), led us to interpret the novelty preference in terms of linguistic maturation. In this section, we discuss the novelty preference in more detail.

We first compare the dialectal children to the non-dialectal children in our study and the results of these dialectal children to the behavior of bivarietal children studied by van Heugten and Johnson (2017). Recall that van Heugten and Johnson (2017) showed that Canadian English children exposed to multiple varieties of English (including foreign-accented speech) only recognized words at the age of 18 months, which is the upper age limit in the dialectal group in our study. There are a number of differences to our study that may have affected the seemingly contradicting direction of the effect, including (a) a remote experiment-controlled vs. a lab-based child-controlled

procedure, and (b) the amount and kind of exposure to the variety tested in the paradigm and to other varieties. Regarding (a), the procedure may certainly influence the results, but it is hard to predict in which way. In a remote setting, there is probably more distraction, which may lead to smaller effect sizes, but hardly to a reversal of the effect. As discussed in section “Discussion,” a different analysis of the looking times, which is closer to a child-controlled procedure, did not change our pattern of results. Furthermore, all of our groups were tested with the App, but there were still differences in preferences. We hence conclude that factors beyond the mere difference in procedure need to account for the difference in findings. We see the most striking differences with regard to (b). However, the actual amount of exposure is difficult to compare. For the bivariate children in van Heugten and Johnson (2017) Canadian English was available 34% of the time. We do not have such an estimate, but we know, from a subsample of the children, that Standard German word forms are frequent in the input of dialectal speech as well (amounting to 66% of the word forms). This difference in input frequency may explain why our dialectal children recognized the word forms earlier than the bivariate Canadian children in van Heugten and Johnson (2017). Beyond the *amount* of input in the variety tested in the word form recognition paradigm [Canadian English in van Heugten and Johnson (2017) and Standard German in our study], which clearly differed in the two studies, the *kind* of input was also different: in van Heugten and Johnson (2017), only one caregiver had an accent different from the one tested in the experiment. It is likely that not many other people speak the same variety, so that exposure to this variety is limited to a single speaker. From the literature on word recognition and word learning, we know that variability (of different sorts) may have beneficial effects on the formation of lexical representations (Singh, 2008; Rost and McMurray, 2009; Höhle et al., 2020). In a study concerned with variability induced by different speakers, Rost and McMurray (2009), for instance, demonstrated that 14-month-old children benefited when novel objects were labeled by different speakers (as compared to single-speaker labeling), cf. Höhle et al. (2020). Singh (2008) compared whether or not stimuli in a familiarization phase showed variability (mixed affect) or not (only positive or negative affect). Their results similarly show that 7.5-month-old children form more specific lexical representations in the high- than in the low-variability condition. The benefits of high variability have so far been documented for variability in the experimental setting (e.g., habituation). In our case, however, speaker and dialectal variability is present in the daily, long-term input that a child receives. This might have indeed boosted the formation of lexical representations (or resulted in different kinds of representations), and in turn, might have very likely led to a novelty preference.

Effects of Dialectal Exposure on Lexical Development

In this section, we briefly reflect on the nature of lexical (and prelexical) representations. While there are a number of studies that have investigated the nature of lexical representations in 20–24-month-old multivarietal children (Flocchia et al., 2012;

Durrant et al., 2015; van der Feest and Johnson, 2016), findings are inconclusive [e.g., Durrant et al. (2015) proposed underspecified representations because the children from the South-West of England recognized both correctly pronounced words and mispronunciations while Flocchia et al. (2012) found that bivariate children did not recognize the words when spoken in the non-rhotic variety of their parents, only in the rhotic variety of the community]. These studies are hard to compare, not least because of the variability in the language varieties, sound contrasts, conditions (same vs. different speaker) and age groups that were tested. The question about the nature of lexical representations in those children remains and goes back to the early stages of development that can be addressed with the familiar word paradigm. The finding that dialectal 12–18-month-old children exhibit looking time differences between Standard German word and nonce-word lists (longer looking times to nonce-word lists) suggests that they recognize Standard German word forms. This novelty preference is already present in the first half of the experiment, showing that dialectal children knew these Standard German word forms before the experiment started. Our results seem to rule out a single storage of the dialectal word form only [as suggested by Flocchia et al. (2012) on the basis of referential word recognition studies]. The next step is thus to test whether dialectal children will exhibit a similar novelty effect as shown for Standard German stimuli when hearing stimuli spoken in their own dialect or in an unfamiliar dialect (with unattested sound alternations).

The novelty preference of dialectal children suggests that they have formed different representations than the non-dialectal children. Currently, we assume that dialectal children link the word forms of the two varieties to each other, either at the prelexical level, where, for example, [u:] is linked to [u̯], or at the conceptual level, where the concept FOOT is linked to [fu:s] and [fūəs]. This would allow them to recognize both Standard German and dialectal forms efficiently. This hypothesis is in line with e.g., Schmale et al. (2011), who showed that “exposure to phonetic variability [during word learning] leads to more robust representations by promoting broader lexical categories” (p. 1105). The additional connections may lead to the observed advantage in processing, exhibited as a novelty preference. There are other studies suggesting different processing mechanisms as a consequence of bilingual input (Meuter and Allport, 1999; Costa and Santesteban, 2004). These studies have shown asymmetric language switching costs in picture naming for L2 speakers, but symmetric switching costs for bilingual speakers. The prediction of more connected representations in dialectal children will be tested in future studies. If our assumption is correct, we predict a novelty preference for dialectal stimuli as well, while stimuli from unfamiliar dialects will not be recognized. For these studies it was relevant to test whether the effects of word type are independent of the segmental nature of the stimuli. In the present study, we compared u-only lists (in which all items contained the stressed vowel [u:]) to u-varied lists (in which half of the items contained [u:] and half contained other vowels). Both types of lists resulted in the same familiarity (or novelty) preferences. These stimuli

hence proved to be well-suited to investigate the nature of representations more closely.

Once more data will be collected with the App, it may also be interesting to differentiate between the dialects the children receive. Currently, most of the dialectal children grow up in the Alemannic dialect area and it is unclear how well the results generalize to other, more conservative, dialects with less transparent phonological mappings between Standard German and dialectal form.

Evaluation of the Remote Testing Using the App

Using the App had a number of advantages compared to traditional laboratory settings: First and foremost, we had access to a larger, more diverse group of participants, in particular from rural areas in which children are exposed to more dialectal forms than the typical child participant tested in the lab. It has to be noted, however, that access to these communities was often only possible through personal contacts who then encouraged their network(s) to participate. This sometimes resulted in participation of children who did not fall into our primary group of interest, given the hypotheses (e.g., older children). Ads in local newspapers and flyers in kindergartens proved to be inefficient for recruitment purposes. Another advantage of the App is that parents do not have to make an appointment to come to the lab. They can freely choose to start the experiment whenever the child is in a good mood. The child likewise benefits from a familiar environment (in contrast to potentially intimidating settings in the lab). Finally, the time investment for most of the drop-outs was minimal, because most of the exclusion criteria were extractable within minutes. The automatic coding of the looking behavior worked extremely well, with a high level of accuracy. The most time-consuming aspect was the manual coding of two of the trials, which took 6–10 min for trained annotators (for one trial of 15 s duration). The manual rating of the dialectal input was done in less than 2 min.

However, home testing leaves researchers with less control over environmental factors. There were cases in which disruptions occurred during testing (e.g., by people talking in the background), which would not have occurred in the lab. Furthermore, the home set-up also assigned more responsibility to the parents, e.g., in reading the instructions, placing the iPad at an appropriate distance and angle, and the restriction not to interfere. Even though parents were explicitly told to participate only once with their child, some parents initiated several attempts (in these cases, we analyzed only the first attempt). The dropout-rate was higher for the older age group (28.6%) than for the younger age group (13.7%); on average it was 19.0% and thus comparable to familiar word recognition studies tested under laboratory conditions (on average 16% in the papers included in Carbajal et al., 2021). iPad-specific reasons for drop-outs were an inadequate position of the tablet so that the child was not in frame, technical issues or loud background noise, but these occurred only very rarely (6 out of 15 dropouts). Furthermore, the experiment-controlled duration of the trials led

to boredom in some children (and to ceiling effects in others), which could have been avoided with a child-controlled set-up. However, a child-controlled set-up requires online-coding of looking times in order to stop trials and initiate new ones. This technical solution was not available at the time of testing. An additional aspect, which typically plays a minor role in the lab, concerned data protection: Many parents were worried about uploading the video data, which prevented some families from participation. Other parents did not have access to iPads, which we solved by lending iPads from the lab to interested parents. Finally, in an attempt to make the use of the App more attractive and to prevent parents from exiting the App before starting the word recognition experiment, the background questionnaires were shorter than those used in the laboratory. This compromise, however, made it hard to capture the input that children received based on the questionnaire alone (e.g., some parents mentioned two languages under the first language field of the App (which we interpreted as bilingual), others mentioned further languages in subsequent language fields of the App). Since we did not ask them to specify in which situations, how often and by whom other languages than German are used, it was hard to specify exclusion criteria. The recordings of two caregivers (which we got from four children) definitely helped to get a better understanding of the child's linguistic environment, but only few parents made use of this option. We do not know whether making this part of the experiment compulsory would prevent them from taking part.

The effect sizes in our App-based experiments were all lower than the average effect size reported in child-controlled familiar word paradigms tested in the lab. Although lower effect sizes may have a number of different causes, we assume that the remote testing with the experiment-controlled stimulus duration may be relevant. Replicating our study with the same group of children and stimuli in the lab will shed light on this issue. In any case, the data suggests that the familiar word effect is robust enough to be replicated with experiment-controlled stimulus duration in home environments. As discussed above (see section "Direction of the Preference"), the reversed effect directions observed across groups is very unlikely to be due to App-based testing, and builds on group differences instead; otherwise, we would have observed a similar pattern in both groups of children.

In future research, a browser version of the experiment could help making the study more widely accessible, even though this might come along with compatibility issues of individual browsers. Furthermore, we plan to test whether the manual coding effort can be further reduced with equally reliable results for the automatic coding. Finally, we are currently testing phonetic fingerprints to distinguish the non-dialectal from the dialectal children (Behrens-Zemek and Braun, 2021).

Home-based testing seems to be a viable option to gather looking-time data of children who grow up with a dialect, which allows us to investigate the development of word forms in populations that hear both Standard German and dialectal forms. The looking-time data indicates that Standard German word forms are recognized by dialectal 12–18-month-old children; the

reversal of the preference (a novelty preference in dialectal as compared to a familiarity preference for non-dialectal children) suggests differences in word form representations, which will have to be investigated in more detail in future studies.

DATA AVAILABILITY STATEMENT

The stimuli and the raw data supporting the conclusion of this article are available on <https://data.mendeley.com/datasets/gf7hsh932v/2>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Konstanz IRB Board. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

BB and KZ-R developed the idea and the design of the study and materials (in collaboration with JR), and the video coding protocol (in collaboration with NC and JK). KZ-R and JR recorded and prepared the experimental stimuli. NC, JK, and KZ-R trained and supervised the annotators for video coding. BB led the statistical analysis and drafted the manuscript. CZ developed the App. JP and BG contributed to the algorithms for automatic video coding. All authors wrote parts and edited the draft.

REFERENCES

- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1080/00273171.2021.1889946
- Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). *The CELEX Lexical Database [CD-ROM]: Linguistic Data Consortium*. Philadelphia, PA: University of Pennsylvania.
- Bayer, J. (1984). COMP in bavarian syntax. *Linguist. Rev.* 3, 209–274.
- Behrens-Zemek, H., and Braun, B. (2021). "Classification of vowels in infant-directed speech as dialectal vs. non-dialectal," in *Paper Presented at the Phonetik und Phonologie im Deutschsprachigen Raum [Phonetics and Phonology in the German-Speaking Area]*, Frankfurt.
- Best, C. T., Tyler, M. D., Gooding, T. N., Orlande, C. B., and Quann, C. A. (2009). Development of phonological constancy: toddlers' perception of native- and Jamaican-accented words. *Psychol. Sci.* 20, 539–542. doi: 10.1111/j.1467-9280.2009.02327.x
- Bornstein, M. H., Hahn, C.-S., Suwalsky, J. T. D., and Haynes, O. M. (2003). "Socioeconomic status and child development: the hollingshead four-factor index of social status and the socioeconomic index of occupations," in *Socioeconomic Status, Parenting, and Child Development*, eds M. H. Bornstein and R. H. Bradley (Mahwah, NJ: Lawrence Erlbaum Associates), 29–82.
- Brandner, E., and Saltzman, M. (2009). Crossing the lake: motion verb constructions in bodensee-alemannic and Swiss German. *Groninger Arbeiten zur germanistischen Linguistik* 48, 81–113.
- Brennan, E. M., Ryan, E. B., and Dawson, W. E. (1975). Scaling of apparent accentedness by magnitude estimation and sensory modality matching. *J. Psychol. Res.* 4, 27–36. doi: 10.1007/BF01066988
- Brugman, H., and Russel, A. (2004). "Annotating multimedia/multi-modal resources with ELAN," in *Paper Presented at the Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*, Lisbon.
- Butler, J., Floccia, C., Goslin, J., and Panneton, R. (2011). Infants' discrimination of familiar and unfamiliar accents in speech. *Infancy* 16, 392–417. doi: 10.1111/j.1532-7078.2010.00050.x
- Carbajal, M. J., Peperkamp, S., and Tsuji, S. (2021). A meta-analysis of infants' word-form recognition. *Infancy* 26, 369–387. doi: 10.1111/inf.12391
- Costa, A., and Santesteban, M. (2004). Lexical access in bilingual speech production: evidence from language switching in highly proficient bilinguals and L2 learners. *J. Mem. Lang.* 50, 491–511. doi: 10.1016/j.jml.2004.02.002
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334.
- Dapretto, M., and Bjork, E. (2000). The development of word retrievabilities in the second year and its relation to early vocabulary growth. *Child Dev.* 71, 635–648. doi: 10.1111/1467-8624.00172
- DeCat, C. (2020). Predicting language proficiency in bilingual children. *Stud. Second Lang. Acquis.* 42, 279–325. doi: 10.1017/s0272263119000597
- DePaolis, R., Keren-Portnoy, T., and Vihman, M. M. (2016). Making sense of infant familiarity and novelty responses to words at lexical onset. *Front. Psychol.* 7:715. doi: 10.3389/fpsyg.2016.00715

FUNDING

The development of the App was funded by an Independent Research Grant of the Institute for Advanced Study for Junior Researchers at the University of Konstanz (awarded to KZ-R).

ACKNOWLEDGMENTS

We thank Carina Haase for recording the stimuli, Monika Lindauer for organizing many of the appointments, for discussion of manuscript and for the provision of literature on socio-economic status, Hendrik Behrens-Zemek, Mirsada Rasidovic Sabanovic, and Naomi Reichmann for manual coding of looking behavior, Moritz Jakob, Justin Hofenbitzer, Sophie Kutscheid, Johanna Schnell, Lena Friedek, Naomi Reichmann, Elena Schweizer, and Friederike Hohl for coding the dialect strength. We further thank Maria Zahner, Beate Zerle and all of the above for help in recruiting, as well as two families for providing us with pictures that were used in the App's instructional video. Also, we are grateful to Christin Beck for programing the script to check the phonotactic probabilities and Achim Kleinmann for extracting the wav-files. We also thank the KinderSchaffenWissen Team (<https://kinderschaffenwissen.eva.mpg.de>) for taking the initiative to promote remote studies in infancy research in Germany and Christoph Schwarze for comments on an earlier version of the manuscript. Finally, we thank all parents and children for taking part in the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.714363/full#supplementary-material>

- Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., et al. (2015). "Long-term recurrent convolutional networks for visual recognition and description," in *Paper Presented at the Computer Vision and Pattern Recognition*, Boston, MA. doi: 10.1109/TPAMI.2016.25919174
- Durrant, S., Delle Luche, C., Cattani, A., and Floccia, C. (2015). Monodialectal and multidialectal infants' representation of familiar words. *J. Child Lang.* 42, 447–465. doi: 10.1017/S0305000914000063
- ELAN (2020). (Version 6.0). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive.
- Ensminger, M. E., and Fortherrgill, K. E. (2003). "A decade of measuring SES: what it tells us and where to go from here," in *Socioeconomic Status, Parenting, and Child Development*, eds M. H. Bornstein and R. H. Bradley (Mahwah, NJ: Lawrence Erlbaum Associates), 13–27.
- Floccia, C., Butler, J., Girard, F., and Goslin, J. (2009). Categorization of regional and foreign accent in 5- to 7-year-old British children. *Int. J. Behav. Dev.* 33, 366–375.
- Floccia, C., Delle Luche, C., Durrant, S., Butler, J., and Goslin, J. (2012). Parent or community: where do 20-month-olds exposed to two accents acquire their representation of words? *Cognition* 124, 95–100. doi: 10.1016/j.cognition.2012.03.011
- Frank, M. C., Braginsky, M., Yurovsky, D., and Marchman, V. A. (2017). Wordbank: an open repository for developmental vocabulary data. *J. Child Lang.* 44, 677–694. doi: 10.1017/S0305000916000209
- Frøta, S., Butler, J., and Vigário, M. (2014). Infants' perception of intonation: is it a statement or a question? *Infancy* 19, 194–213. doi: 10.1111/inf.12037
- Grewendorf, G., and Weiß, H. (eds) (2014). *Bavarian Syntax: Contributions to the Theory of Syntax*. Amsterdam: John Benjamins.
- Grondelaers, S., van Hout, R., and van der Harst, S. (2015). Subjective accent strength perceptions are not only a function of objective accent strength. Evidence from Netherlandic Standard Dutch. *Speech Commun.* 74, 1–11.
- Hallé, P. A., and Boysson-Bardies, B. D. (1994). Emergence of an early receptive lexicon: infants' recognition of words. *Infant Behav. Dev.* 17, 119–129. doi: 10.1016/0163-6383(94)90047-7
- Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., and Geyken, A. (2011). dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung [dlexDB - a lexical database for psychological and linguistic research]. *Psychol. Run.* 62, 10–20. doi: 10.1026/0033-3042/a000029
- Hoff, E., Laursen, B., and Tardif, T. (2002). "Socioeconomic status and parenting," in *Handbook of Parenting: Biology and Ecology of Parenting*, Vol. 2, ed. M. H. Bornstein (Mahwah, NJ: Lawrence Erlbaum Associates), 231–252.
- Höhle, B., Fritzsche, T., Mess, K., Philipp, M., and Gafos, A. (2020). Only the right noise? Effects of phonetic and visual input variability on 14-month-olds' minimal pair word learning. *Dev. Sci.* 23:e12950. doi: 10.1111/desc.12950
- Houston-Price, C., and Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant Child Dev.* 13, 341–348.
- Hunter, M. A., and Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Adv. Infancy Res.* 5, 69–95.
- Johnson, E. K. (2016). Constructing a proto-lexicon: an integrative view of infant language development. *Annu. Rev. Linguis.* 2, 391–412. doi: 10.1146/annurev-linguistics-011415-040616
- Kalashnikova, M., Goswami, U., and Burnham, D. (2016). Delayed development of phonological constancy in toddlers at family risk for dyslexia. *Infant Behav. Dev.* 57:101327. doi: 10.1016/j.infbeh.2019.101327
- Kleber, F. (2020). Complementary length in vowel–consonant sequences: acoustic and perceptual evidence for a sound change in progress in Bavarian German. *J. Int. Phonet. Assoc.* 50, 1–22. doi: 10.1017/s0025100317000238
- Kohler, K. (1990). "Segmental reduction in connected speech in German: phonological facts and phonetic explanations," in *Speech Production and Speech Modelling*, eds W. Hardcastle and A. Marchal (Dordrecht: Kluwer), 69–92. doi: 10.1007/978-94-009-2037-8_4
- Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.
- Levy, H., Konieczny, L., and Hanulíková, A. (2019). Processing of unfamiliar accents in monolingual and bilingual children. Effects of type and amount of accent experience. *J. Child Lang.* 46, 368–392. doi: 10.1017/S030500091800051X
- Meuter, R. F. I., and Allport, A. (1999). Bilingual language switching in naming: asymmetrical costs of language selection. *J. Mem. Lang.* 40, 25–40. doi: 10.1006/jmla.1998.2602
- Morey, R. D., and Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs (R package version 0.9.12-4.2)*. Retrieved online at: <https://CRAN.R-project.org/package=BayesFactor> (accessed August 10, 2021).
- Munske, H. H. (2015). "Der Bayerische Sprachatlas (BSA) [The Bavarian language atlas]," in *Regionale Variation des Deutschen*, eds R. Kehrein, A. Lameli, and S. Rabanus (Berlin: Walter de Gruyter), 1–28.
- Noble, K. G., Norman, M. F., and Farah, M. J. (2005). Neurocognitive correlates of socioeconomic status in kindergarten children. *Dev. Sci.* 8, 74–87. doi: 10.1111/j.1467-7687.2005.00394.x
- Pollock, S., and Nazzi, T. (2015). Consonant/vowel asymmetry in early word form recognition. *J. Exp. Child Psychol.* 131, 135–148. doi: 10.1016/j.jecp.2014.11.011
- Porretta, V., Tucker, B. V., and Järvikivi, J. (2016). The influence of gradient accentedness and listener experience on word recognition. *J. Phonet.* 58, 1–21. doi: 10.1016/j.wocn.2016.05.006
- Rost, G. C., and McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Dev. Sci.* 12, 339–349. doi: 10.1111/j.1467-7687.2008.00786.x
- Ryan, E. B. (1973). "Subjective reactions toward accented speech," in *Language Attitudes: Current Trends and Prospects*, eds R. W. Shuy and R. W. Fasold (Washington, DC: Georgetown University Press), 60–73.
- Schmale, R., Cristia, A., Seidl, A., and Johnson, E. K. (2010). Developmental changes in infants' ability to cope with dialect variation in word recognition. *Infancy* 15, 650–662. doi: 10.1111/j.1532-7078.2010.0032.x
- Schmale, R., Hollich, G., and Seidl, A. (2011). Contending with foreign accent in early word learning. *J. Child Lang.* 38, 1096–1108. doi: 10.1017/S0305000910000619
- Schwarz, C. (2014). Conservative and innovative dialect areas. *Taal Tongval* 66, 65–83. doi: 10.5117/tet2014.1.schw
- Siebenhaar, B., and Wyler, A. (1997). *Dialekt und Hochsprache in der Deutschsprachigen Schweiz [Dialect and Standard in the German-speaking part of Switzerland]*. Zürich: Pro Helvetia, Schweizer Kulturstiftung.
- Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition* 106, 833–870. doi: 10.1016/j.cognition.2007.05.002
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: a meta-analytic review of research. *Rev. Educ. Res.* 75, 417–453. doi: 10.3102/00346543075003417
- Stöten, K., and Engstrand, O. (2003). Effects of perceived age on perceived dialect strength: a listening test using manipulations of speaking rate and F0. *PHONUM* 9, 29–32.
- Swingle, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Dev. Sci.* 8, 432–443. doi: 10.1111/j.1467-7687.2005.00432.x
- Szgun, G., Stumper, B., and Schramm, S. A. (2009). *Fragebogen zur Frühkindlichen Sprachentwicklung (FRAKIS) und FRAKIS-K (Kurzform)*. [Questionnaire on the early child language development (FRAKIS) and FRANKIS-K (short form)]. Frankfurt: Pearson Assessment.
- Thiessen, E. D., Hill, E. A., and Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy* 7, 53–71. doi: 10.1207/s15327078in0701_5
- van Bezooijen, R., and van Hout, R. (1985). Accentedness ratings and phonological variables as measures of variation in pronunciation. *Lang. Speech* 28, 129–142. doi: 10.1177/002383098502800203
- van der Feest, S. V. H., and Johnson, E. K. (2016). Input-driven differences in toddlers' perception of a disappearing phonological contrast. *Lang. Acquis.* 23:150511085813008. doi: 10.1080/10489223.2015.1047096
- van Heugten, M., and Johnson, E. K. (2014). Learning to contend with accents in infancy: benefits of brief speaker exposure. *J. Exp. Psychol.* 143, 340–350. doi: 10.1037/a0032192
- van Heugten, M., and Johnson, E. K. (2017). Input matters: multi-accent language exposure affects word form recognition in infancy. *J. Acous. Soc. Am.* 142:EL196. doi: 10.1121/1.4997604

- van Heugten, M., Paquette-Smith, M., Krieger, D. R., and Johnson, E. K. (2018). Infants' recognition of foreign-accented words: flexible yet precise signal-to-word mapping strategies. *J. Mem. Lang.* 100, 51–60. doi: 10.1016/j.jml.2018.01.003
- Vihman, M. M., and Majorano, M. (2017). The role of geminates in infants' early word production and word-form recognition. *J. Child Lang.* 44, 158–184. doi: 10.1017/S0305000915000793
- Vihman, M. M., Nakai, S., DePaolis, R., and Hallé, P. (2004). The role of accentual pattern in early lexical representation. *J. Mem. Lang.* 50, 336–353. doi: 10.1016/j.jml.2003.11.004
- Vihman, M. M., Thierry, G., Lum, J., Keren-Portnoy, T., and Martin, P. (2007). Onset of word-form recognition in English, Welsh, and English-Welsh bilingual infants. *Appl. Psycholinguist.* 28, 475–493. doi: 10.1017/s0142716407070269
- Vitevitch, M. S., and Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behav. Res. Methods Instrum. Comput.* 36, 481–487. doi: 10.3758/bf03195594
- Westermann, G., and Mani, N. (eds) (2018). *Early Word Learning (Current Issues in Developmental Psychology)*. New York, NY: Routledge.
- White, K. S. (2018). "Listening to (and listening through) variability during word learning," in *Early Word Learning*, eds G. Westermann and N. Mani (Routledge: Taylor & Francis), 83–95. doi: 10.4324/9781315730974-7
- Zahner, K., Jakob, M., Lindauer, M., and Braun, B. (2021). "Child-directed-speech is not affected by recording setting: preliminary results on Southern German and Swiss German," in *Paper Presented at the Phonetics and Phonology in Europe* (Lisbon).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Braun, Czeke, Rimpler, Zinn, Probst, Goldlücke, Kretschmer and Zahner-Ritter. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Tracking Infant Development With a Smartphone: A Practical Guide to the Experience Sampling Method

Marion I. van den Heuvel^{1*}, Anne Bülow^{2,3}, Vera E. Heininga^{4,5}, Elisabeth L. de Moor⁶,
Loes H. C. Janssen⁷, Mariek Vanden Abeele⁸ and Myrthe G. B. M. Boekhorst^{1,9}

¹Department of Cognitive Neuropsychology, Tilburg University, Tilburg, Netherlands, ²Department of Psychology Education & Child Studies, Erasmus University Rotterdam, Rotterdam, Netherlands, ³Department of Developmental Psychology, Tilburg University, Tilburg, Netherlands, ⁴Department of Developmental Psychology, Groningen University, Groningen, Netherlands, ⁵Department of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium, ⁶Department of Youth and Family, Utrecht University, Utrecht, Netherlands, ⁷Department of Clinical Psychology, Leiden University, Leiden, Netherlands, ⁸imec-mict-UGent, Department of Communication Sciences, Ghent University, Ghent, Belgium, ⁹Department of Medical and Clinical Psychology, Tilburg University, Tilburg, Netherlands

OPEN ACCESS

Edited by:

Natasha Kirkham,
Birkbeck, University of London,
United Kingdom

Reviewed by:

Randy Corpuz,
University of Massachusetts Boston,
United States
Erica Neri,
University of Bologna, Italy

*Correspondence:

Marion I. van den Heuvel
m.i.vdnheuvel@tilburguniversity.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 April 2021

Accepted: 03 November 2021

Published: 06 December 2021

Citation:

van den Heuvel MI, Bülow A,
Heininga VE, de Moor EL,
Janssen LHC, Vanden Abeele M and
Boekhorst MGBM (2021) Tracking
Infant Development With a
Smartphone: A Practical Guide to the
Experience Sampling Method.
Front. Psychol. 12:703743.
doi: 10.3389/fpsyg.2021.703743

The COVID-19 pandemic has forced developmental researchers to rethink their traditional research practices. The growing need to study infant development at a distance has shifted our research paradigm to online and digital monitoring of infants and families, using electronic devices, such as smartphones. In this practical guide, we introduce the Experience Sampling Method (ESM) – a research method to collect data, in the moment, on multiple occasions over time – for examining infant development at a distance. ESM is highly suited for assessing dynamic processes of infant development and family dynamics, such as parent-infant interactions and parenting practices. It can also be used to track highly fluctuating family dynamics (e.g., infant and parental mood or behavior) and routines (e.g., activity levels and feeding practices). The aim of the current paper was to provide an overview by explaining what ESM is and for what types of research ESM is best suited. Next, we provide a brief step-by-step guide on how to start and run an ESM study, including preregistration, development of a questionnaire, using wearables and other hardware, planning and design considerations, and examples of possible analysis techniques. Finally, we discuss common pitfalls of ESM research and how to avoid them.

Keywords: experience sampling method/ecological momentary assessment, infant development, longitudinal data, ambulatory monitor, infancy

INTRODUCTION

The COVID-19 pandemic made traditional observational and experimental methods unavailable due to closing labs and the introduction of strict rules regarding physical contact. This forced infant developmental researchers to rethink their traditional research practices by using novel paradigms and electronic devices. One way of studying infant development without physical contact is to use the Experience Sampling Method (ESM) – also called Ecological Momentary Assessment (EMA), Ambulatory Assessment (AA), or Mobile Experience Sampling Method (MESM, m-ESM) – a research method to collect data, in the moment, on multiple occasions

over time (Csikszentmihalyi and Larson, 1987; Stone and Shiffman, 1994). In infant research, this often takes the form of the parent filling out questionnaires about their infant's behavior and mood, tracking their infant's sleep or activity patterns, and/or reporting on their own mood, thoughts, practices, and behaviors.

The first ESM-type studies, from the 1980s onwards, have provided unique insights into daily life processes by using paper-and-pencil questionnaires or "diaries." Infant researchers mainly used diary measures in which mothers and fathers kept track of their infant's activities, such as sleep (St James-Roberts and Plewis, 1996) or cry and fuss behavior (Barr et al., 1988; Fujiwara et al., 2011). To date, diary work is still used to measure infant sleep and crying behavior and is often even considered more valid than using one-time questionnaires (Teti et al., 2010; Fujiwara et al., 2011; Hechler et al., 2018; Bacaro et al., 2020). Despite the strong tradition in diary work in infant research and the large overlap in methods between diary work and ESM research, the use of ESM remains sparse in infant research. With the COVID-19-infused shift in research practices, there may be momentum to welcome ESM to the skillset of infant researchers and use it to innovate the field.

Over the years, researchers have provided systematic guides and overviews of ESM methodology to discuss the various steps that are important to consider when setting up an ESM study (e.g., sample size calculation, software, implementation, and data analysis) (Christensen et al., 2003; Uy et al., 2009). Even though these studies have provided an important outline of ESM research, they were conducted prior to the regular use of the smartphone (e.g., using computerized methods, Personal Data Assistants (PDAs), and paper-and-pencil methods). Over the recent years, the daily use of smartphones has become more common. As a result, many ESM applications and software for smartphones have been developed. Smartphones are therefore a major step forward in ESM research, allowing for more accurate "in-the-moment" collection of data. Thus, we build on this previous literature by specifically catering to developmental researchers and by providing a more grounded overview and step-by-step guide of ESM research for smartphone monitoring. The aim of this practical guide was to provide infant researchers with an introduction to ESM research for studying infant development and to lower the potential barrier for including this technique into their skillset. We start this practical guide by explaining what ESM entails and by discussing advantages and disadvantages for using ESM in infant research. Next, we provide a brief step-by-step guide on how to design and run an ESM study, including preregistration, selection/development of questionnaires, design planning, running the study, and analysis techniques. Finally, we discuss common pitfalls of ESM research and how to avoid them.

ESM IN INFANT RESEARCH

What Is ESM Research?

With ESM, participants fill out micro-surveys several times a day for several consecutive days. ESM captures "Life as it is lived," by assessing participants' cognitions, emotions, and

experiences several times a day in the context of daily life (Bolger et al., 2003; Scollon et al., 2003). Note that a diary study could also be considered as ESM, when it asks parents to make diary entries multiple times a day, such as every half hour or after each nap of their infant (e.g., Barr et al., 1988). However, many diary studies require participants to recall events, feelings, and/or thoughts only once a day (i.e., daily diary), for example, before going to bed. Additionally, ESM research nowadays is often conducted with the use of a smartphone or other electronic device and can be accompanied by other technologies (e.g., activity watch, heart rate sensor, and/or beacon).

Infant Research Using ESM

Experience sampling method studies that investigate infant development (up to the age of 24 months) and family dynamics can include parental-related measures (e.g., maternal mood and feeding practices), infant-related measures (e.g., crying and sleep), or a combination of both. Only very few studies so far have incorporated ESM to study these measures. One example is a study by Sawada et al. (2015), that included ESM to assess daily maternal reports of infant fussing and crying at 12 months postpartum. They studied associations between maternal stress (conceptualized as "felt security") at 6 months postpartum and infant fussing and crying at 12 months postpartum in mother-infant dyads with a healthy infant ($N = 93$) and dyads with infants who have a medical problem ($N = 42$). Infant fussing and crying was measured with ESM by paging the mothers with a personal digital assistant (PDA) three times a day for 7 days. Among dyads with an infant born with medical problems, higher felt security of the mother predicted decreased fussing and crying of the infant, but not among dyads with a healthy infant. Another study used ESM to track soothing behaviors of 157 mothers after an intervention that focused on reducing the use of feeding to soothe an infant (Adams et al., 2019). The mothers reported on infant fussing and crying with 4-h intervals, between 10 AM and 10 PM, and filled out a morning and evening questionnaire. They found that the parenting intervention was successful in reducing the use of food to soothe and increased the use of alternative soothing strategies in response to infant fussiness. With this ESM technique, the authors were able to gather detailed, ecologically valid data on mothers' soothing techniques, right after soothing took place.

Another research group recently used ESM to follow pregnant mothers with a substance use disorder to examine momentary fluctuations in posttraumatic stress disorder (PTSD) symptoms, prenatal bonding, and substance craving (Sanjuan et al., 2020). The pregnant women filled out three 5- to 10-min questionnaires for 28 days asking about PTSD symptoms, prenatal bonding (i.e., the quality of maternal affective experience and intensity of preoccupation with fetus), and substance cravings. Results showed that higher momentary ratings of PTSD symptoms in these mothers were associated with lower quality (but not intensity of preoccupation) of prenatal bonding, which in turn was associated with greater craving for substances. Next, another study used ESM to capture maternal experiences and emotions

in the context of real-world, day-to-day parenting challenges (Hajal et al., 2019). For this study, mothers (N=55) were interviewed over the phone four times a day for 6 days about their momentary emotions, motivational states, and parenting behaviors. They found that maternal reported momentary emotions were more consistently associated with momentary motivational states (i.e., desire to approach/engage and avoid/disengage with their infant) than reported behaviors and underscore the importance of momentary emotions in studying family dynamics.

Finally, a recent study used ESM to assess the relationship between testosterone and time fathers invested in their infants (Corpuz et al., 2021). In this study, testosterone was measured in first-time fathers (N=225) during their transition into parenthood (pregnancy, 3 months postpartum, and 9–10 months postpartum). For the ESM part of the study, the authors assessed the time invested in direct infant care. To measure this, participating fathers received eight text messages for 6 days that they were not working. The study found a relationship between accelerated testosterone rebound (increase) and less time spent with their infant. There was also a positive association between testosterone rebound and the quality of care fathers showed (not measured using ESM).

In sum, while ESM research in infancy is sparse, the summarized studies show the feasibility of using ESM in pregnant women and early postpartum mothers (even when suffering from PTSD and drug abuse) to assess dynamic and momentary processes regarding infant and maternal mood, as well as infant-parent interactions and father involvement in infant care. However, many research questions in infant development on family dynamics remain unexplored. ESM could add potential new insights given its advantages over experimental and questionnaire research, allowing for momentary assessment of the infant-parent dynamics and frequent fluctuations of mood and behavior.

Advantages of ESM for Infant Research

The most notable advantage of ESM is that it enables researchers to capture data “*in situ*” (Naab et al., 2019). With ESM, researchers can gain information not only about *content*, but also its *context* (Hektner and Csikszentmihalyi, 2002). Because ESM sheds light on the “situatedness” of human experience, it is a highly suitable method for studying context-dependent processes that occur during infant development (Bamberger, 2016). For instance, researchers can assess whether an infant is happy or sad and link that mood to contextual information, such as their sleep quality or maternal mental health. To date, such analyses have not yet been conducted in infancy research. For an example in adolescent research, see Kim et al. (2018).

A second advantage of ESM is that it offers a possibility to assess complex phenomena, such as infant development and family functioning in an *ecologically valid way* (Trull and Ebner-Priemer, 2009) and with reduced recall bias (Schwarz, 2007). Recall bias refers to the bias that arises when people retrospectively report on behavior, emotions, or cognitions,

as is typical in cross-sectional or retrospective longitudinal survey research. ESM reduces recall bias because it requires no or a very limited retrospective recall (Scollon et al., 2009). An alternative to reduce recall bias in self-reports would be to directly observe parents and children in lab settings. Such observations are costly, however, limited in time, do not give access to mental states or cognitions, and lack ecological validity. The latter may be solved by observing in the home or in public settings, but this requires extensive consideration of research ethics, and such observations may also influence the ecological validity as parents behave differently when being knowingly observed (Vanden Abeele et al., 2020).

Third, because ESM enables multiple assessments over a relatively short time span, researchers can collect data about phenomena that are potentially short-lived or transient in nature and, thus, allows researchers to *capture the dynamic nature of events* and shed light on how events unfold in everyday life (Hektner and Csikszentmihalyi, 2002). These events may concern one individual; for instance, researchers can explore whether an infant's lack of sleep predicts subsequent crying behavior. Additionally, ESM is also suited to study family dynamics (Larson et al., 1996; Repetti et al., 2015; Bamberger, 2016). Especially relevant in this regard are studies that collect dyadic data (e.g., parent-infant data), for instance, to assess whether there are transmission effects in parents' and infants' emotional states.

A fourth advantage of ESM is that the examination of temporal patterns offers the possibility to investigate not only between-person, but also *within-person* (or *within-family*) *processes* (Scollon et al., 2003; Keijsers and van Roekel, 2018). This is of great importance in the study of infant development. It is thinkable that group-level associations between parenting behaviors and child development do not uphold when examining within-family associations. For instance, while there may be a positive within-person association between infant sleep duration and infant mood, the between-person association may be absent or even reversed. In other words, it is reasonable to expect that an infant that sleeps longer will be more rested and thus have a better mood (within-person effect). However, it may not be the case that infants that (on average) sleep longer are generally more happy (between-person effect).

A fifth advantage of ESM is that infant development can be studied solely at a distance. This research method utilizes data collection through smartphone, allowing for complete digital monitoring. This is especially important during the COVID-19 pandemic, when observational and experimental research methods were unable to be used due to lockdown regulations (e.g., closed labs and contact restrictions). Apart from during the COVID-19 pandemic, it is also important to mention that research from a distance can be of added value for infant research in general. ESM research would limit travel time, which would be advantageous for parents living further away and infants in general. It can be incorporated in the daily life of parents with young infants, taking parenting schedules into consideration.

Disadvantages of ESM for Infant Research

Assessing parent or infant behavior several times a day also has several disadvantages. It can be *burdensome for parents* (Eisele et al., 2020). With ESM, parents usually receive multiple short questionnaires per day. It is not uncommon for ESM studies, for instance, to administer even up to 8–15 short questionnaires per day (e.g., van Roekel et al., 2019; Dietvorst et al., 2021). Although completing each questionnaire usually takes no more than a few minutes, participation in an ESM study can be demanding for young parents, both in terms of the total time investment and in terms of the (cognitive) resources that are needed to always be able to respond to the questionnaires in a timely and qualitative manner (Sonnenberg et al., 2012).

Because of the burden associated with ESM studies, another significant disadvantage of ESM is *self-selection bias*: Parents who can better cope with the burden of the ESM may be overrepresented in the sample, and this may lead to an overrepresentation of participants in the sample with relevant characteristics, such as greater motivation (Scollon et al., 2009). This disadvantage can be highly relevant for infant research. For parents of young children who already struggle to manage the day-to-day organization of their household or those parents who experience heightened levels of psychopathology, participation in an ESM study may be especially burdensome. This becomes specifically problematic when inclusion of specific subgroups is of interest to the researchers.

A third disadvantage of ESM is potential *bias in the responses*. Since ESM requires the parent to have access to a device (usually their smartphone) to answer questions, not all activities are easy to capture with this technique. For instance, when a parent goes swimming with their infant, they will probably not check their phone, and therefore, an ESM study that focuses on parent–child quality time may miss instances of such activities. Furthermore, in specific subgroups of individuals, such as parents with mental illness, this bias may also play a role. For example, in their guide, Palmier-Claus et al. (2011) mentioned that altered sleep patterns (e.g., unusual waking/sleeping times) are important to consider in individuals with mental illness, as they may therefore not always be able to respond to the ESM triggers. Additionally, by its very nature, ESM will interrupt parents in their daily activities. Especially in the context of studies focusing on infant development, inducing such interruptions for the sake of scientific research comes with ethical considerations. Recent studies showed that mobile device use makes parents up to five times less responsive to young children's bids for attention (Vanden Abeele et al., 2020). Hence, researchers carry a responsibility when asking parents to participate in an ESM study.

PRACTICAL STEP-BY-STEP GUIDE

An overview of the different steps that are important to consider when planning and running an ESM study is provided in **Figure 1**.

Preregistration of an ESM Study

Preregistration of ESM studies is highly recommended. Preregistration is a specification of the research plans in advance of the study and prevents the unintentional use of the same data to both generate and test a hypothesis. Separating exploratory (i.e., hypothesis generating) from confirmatory research (i.e., hypothesis testing) improves the quality and transparency of research. Preregistration of an ESM study has many of the same components as other (laboratory-based) studies of infant development, but also include additional elements, such as sampling scheme, trigger logic, and monitoring strategies (for explanation, see section *planning and programming an ESM study*). In a recent paper by Kirtley et al. (2021), the process of preregistering an ESM study is described and an open-access template is provided.

Developing an ESM Questionnaire

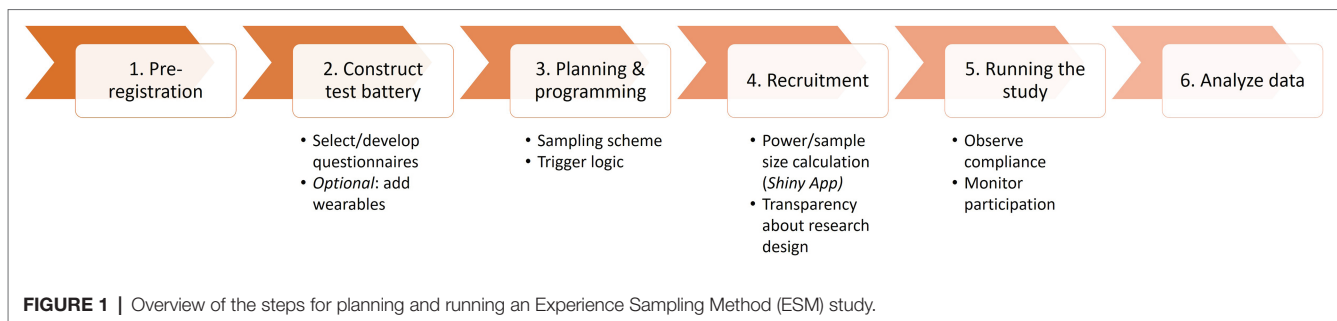
The next step is to create a test battery for the ESM study by selecting instruments. It is important that the instruments used in an ESM study cater to the format of micro-surveys that are administered several times a day and that they target the momentary experiences of participants. In contrast to most single assessment questionnaires, ESM items should accommodate momentary states, making it clear that the question is regarding experiences “*right now*” or refers to the time *in between* assessments (Myin-Germeyns et al., 2018). Some studies have addressed these aspects by adjusting existing questionnaires (e.g., shortening or rewording) that are commonly used for the retrospective design. Nonetheless, different types of questions can be applied to ESM research. In their paper, van Berckel et al. (2017) draw on common ESM question types, addressing the usage and challenges of different types (e.g., checkboxes, text field, sliders, Likert scale, and photos). ESM questions can also be organized in a branching structure to decrease the participants' burden (e.g., using a checklist with multiple answer options and only show the follow-up Likert scale questions for the items selected by the participant). The ESM item repository¹ can facilitate in construction of an ESM-friendly test battery.

To date, there are very few items/questionnaires available that are tailored to infant development and family dynamics and practices, such as feeding practices, sleeping, and mother–infant interaction. An example of an item suitable to assess infant mood within an ESM design could be: “how would you rate your child's mood” (0 = very fussy to 5 = very happy) (Mindell and Lee, 2015). Furthermore, very recently, an ESM questionnaire was developed by the first and last authors (MvdH and MB) to measure maternal baby-related anxiety, the “Baby-related Anxiety and Behavior Inventory (BABI).” The questionnaire and development protocol can be downloaded at OSF (Boekhorst et al., 2021, February 18²).

When developing an ESM questionnaire battery, it is essential that the participants' burden is also taken into consideration

¹<https://esमितemrepositoryinfo.com/>

²<https://osf.io/3tznbl/>



(Consolvo and Walker, 2003). As addressed by Myin-Germeys et al. (2018), an ESM questionnaire should only take 2 min or less to complete and shorter questionnaires lead to higher compliance, especially considering the already busy schedule of young parents raising an infant (see section compliance). Furthermore, it is also important to make sure that the constructs measured are variable (enough) so participants' responses to the recurring questionnaires are not always the same. For instance, ask about behaviors, thoughts, and mood states, instead of more static features, such as opinions, knowledge, and personality. Finally, it is also advised to pilot the questionnaire to estimate psychometric properties, the time it takes, and ask for feedback on feasibility, the likability of the questionnaire, and parents' personal experience.

While this article mainly focuses on ESM as a quantitative research method, it is relevant to note that there are also alternative, more qualitative, approaches to ESM. Kaufmann and Peil (2019), for instance, developed the Mobile Instant Messaging Interview (MIMI) method, in which they use a mobile messenger, such as WhatsApp, to interact with participants at different time points during the day. With the MIMI technique, the researchers, for instance, start a conversation over WhatsApp asking whether the participant is using any social media at that moment by asking the participant to send a description, photo, and/or video. Another example is a recent study by Cho and Ilari (2021), that investigated the association between child mood, music, and parenting during the COVID-19 pandemic. The authors shared music videos with the mothers (to play for their child) and received video and photo materials back from the participants. Adding qualitative items can be valuable in the context of infant development research, as they enable researchers to capture data that is rich in nature. A major drawback of qualitative measures, however, is the time investment of the participant and, in some cases (e.g., sending photos), privacy concerns.

Using Smartphones

The daily use of smartphones has become more common, also allowing researchers to use them to collect data for ESM data collection. As a result, applications have been specifically designed for ESM studies using smartphones. The use of smartphones for ESM research has several advantages. Using smartphones allows for a more efficient implementation of the questionnaires that are completely online (e.g., without needing to use any

paper-pencil questionnaires), the administration is more sufficient for researchers, data collection, and completion is more time-efficient, and ESM applications send automated triggers (e.g., at random, scheduled) to the participants' smartphone, which in turn can help with compliance (Ainsworth et al., 2013; Thomas and Azmitia, 2015; Stieger and Reips, 2019). Using smartphones for ESM data collection are also efficient for infant researchers as parents receive an automated reminder on their smartphone without needing to schedule questionnaire completion themselves, which may be more difficult during parenting practices of an infant (e.g., caretaking, play, feeding).

As the number of ESM studies grow steadily, new mobile applications and software packages are being developed, with a wide range in technical features and pricing. Researchers can consider several things when selecting the ESM software or applications. For example, is working with an application *via* smartphone necessary or would setting up a study in a web tool be sufficient? The latter would enable participants to use multiple devices to answer surveys or include participants without smartphones. However, participants would not be able to answer surveys when they are offline and the study cannot collect additional passive data. Systems also differ in customizability. In some Open Source cases, it is possible to customize the platform entirely, but this might take getting used to the system. When using a mobile application, it is important to consider the possible operating systems. Some only work on Android and others also partly on iOS, but there are also applications who work properly on both operating systems. Although most companies provide a limited free trial or plan, they often use different ways of payment (e.g., payments for a certain amount of time, or payment per participant). Various Open Source platforms are free for all users, while others provide additional licenses for extra features.

Nonetheless, it is important to note that ESM applications are developing rapidly, suggesting that personal research on the possibilities and costs of different applications is essential. From personal experience, companies are usually open to suggestions for adding features to expand their applications or packages. This does not always have to result in paying additional costs, depending on the amount of similar suggestions they receive.

Using Wearables and Other Hardware

Traditionally, ESM studies rely on participants answering questionnaires, but the use of smartphones enables researchers

to also passively collect data on, for instance, health and social activities (i.e., location, accelerometer activity, device usage data, and microphones). Moreover, some applications even allow a seamless integration with wearables for tracking activity, sleep, physiological arousal, or location (Trull and Ebner-Priemer, 2009). Therefore, the next step is to decide whether or not to include wearables or other hardware to the ESM study. Here, we discuss two types of hardware that can be especially useful for infancy research: activity trackers and beacons.

Activity Trackers

To get more insight into, for instance, infant sleeping behavior (i.e., amount of sleep, quality of sleep), it is possible to include an actigraphy monitor (i.e., ActiGraph and Fitbit) in an ESM study. This additional information can help to answer validation questions, such as whether the measured sleep and nap time by the ActiGraph corresponds to maternal self-report (Tikotzky and Sadeh, 2009). Infants typically wear the (mini) ActiGraph on their ankle while they sleep. Such devices can also be used to decrease the burden of asking parents to report on infant sleeping patterns by measuring it (with even more precision) directly.

Beacons

While ESM studies with self-report items of questionnaires can already contribute to a better understanding of the parent-infant relation, gaining even more insight by using an unobtrusive measure to assess actual parent-child proximity in daily life seems promising. Asking parents about the amount of time they spend with their infant can still be biased or influenced by nonresponse. Instead, fine-grained information on parent-infant proximity (i.e., distance in meters) can be obtained when using Bluetooth and Bluetooth beacon devices. When the beacon is placed in the crib or sewn into clothing, researchers can measure the exact times that a mother is close to her infant. Researchers can configure the beacon devices with specific settings (i.e., time interval of checking for other beacons, appropriate distance between beacons) and give a beacon device to each member of a family (i.e., mother, father, and infant) to study family dynamics. Another option is to trigger a questionnaire or another form of data collection, such as audio recording, with beacon information. A very recent example of this is the pilot study by Salo et al. (2021), in which the researchers developed clothing for infants with technology in it (i.e., a beacon and audio recorder) – called the “TotTag.” The audio recorder was triggered and started recording when the mother (who was also wearing a beacon) was close to the infant.

Planning and Programming an ESM Study

Naturally, the design of the ESM study should align with the aims and research questions. When setting up an ESM study, several factors should be taken into consideration, such as the type of sampling and intervals between assessments. Sampling can be time-based, meaning that participants will receive a

notification on fixed or random times as programmed by the researcher. Assessments can also be event-based. In this case, participants have to initiate the completion of an assessment after an event has occurred (such as after feeding or infant naps). Event-based assessment can also be triggered automatically, for instance, when a participant is at a certain location (GPS-based or beacon, see above at *beacons*; e.g., when parents are at home *with* the infant or when they are at work *without* the infant). Importantly, for measuring social situations, such as interactions, the time-based and event-based sampling designs lead to comparable data quality (Himmelstein et al., 2019). After deciding whether event-based and/or trigger-based works best for the purpose of a study, the next two important factors to consider are the sampling scheme and the trigger logic.

Sampling Scheme

When using time-based sampling, it is important to consider issues, such as the time window in which the sampling will occur. When the study aim is regarding parent-infant activities, it is important to take into account whether parents go to work or are otherwise not with their infant (e.g., daycare), but when examining mood over time, it is important to assess mood several times throughout the day since mood fluctuates constantly. When sampling during the night, participants will likely miss several assessments (and it can be very burdensome). When interested in nighttime activities and/or interactions, it is advised to very carefully inform participants and/or ask about nighttime in the morning (e.g., how well did you sleep? and How did the infant sleep?). In some studies, it might be convenient to personalize the sampling scheme to increase compliance, while in other studies, this could affect the outcomes (e.g., data on parental stress/mood during the day might remain unnoticed or inaccurate if questionnaires are only received when it is better suited for the parents, such as during infant naptime or during a day off). Important to note is that not all software applications provide the opportunity to personalize schemes. Additionally, for some analyses, the time interval between the measurements needs to be equal. Therefore, planning analyses before designing the study is advised.

Trigger Logic

Most software applications allow for decisions on a trigger logic, for example, when the questionnaires are prompted and for how long participants can respond (e.g., time-based: trigger questionnaire between 9 and 10:A.M, event-based: trigger when beacon signals contact with infant). Participants can also be reminded to fill out unanswered questionnaires by sending them notifications (most applications allow this). Notifications can be triggered with a logic, for instance after 30 min of not completing the questionnaire. Usually, questionnaires will be inaccessible after a certain time (expired) or when the next questionnaire is prompted.

Recruiting Parents for Your ESM Study

Before recruiting parents for the ESM study, it is advised to calculate an ideal sample size given the analysis plan and

desired power. While power analyses are often used to inform sample size planning in general (Cohen, 1988), they are not yet well-established in ESM research. This is mainly because the multilevel structure of the data (i.e., many nested measurements per participant) makes power calculations challenging (de Jong et al., 2010; Bolger, 2011). Recently, however, LaFit et al. (2021) published a new “*Shiny App*” that helps to perform simulation-based power analyses. Still, pilot data are necessary to be able to perform such a power analysis, to extract estimates for the parameter values (e.g., as explained here:³).

When recruiting participants, it is important to be transparent about the time investment of ESM research and clearly explain *why* it is helpful to the study to collect data at multiple times during the day rather than just once as is the case in traditional retrospective research. When failing to explain why parents need to fill out the same questions “over and over again,” they may dropout or become frustrated. To increase their compliance, it is also important to accommodate their schedules and talk about how they can fit the ESM into their daily lives. Depending on the study design, it can help to offer multiple starting dates to pick from, suiting their schedule best, instead of one start date for all participants.

Running an ESM Study

Every researcher wants high-quality data. In ESM research, one can mostly focus on compliance – the percentage of answered questionnaires – as a data quality indicator. Only if compliance is high, one is able to generalize the answers to daily life and data analysis are sufficiently powered. In most studies, participants answer on average 50–95% of all ESM surveys (Wen et al., 2017; van Roekel et al., 2019; Williams et al., 2021). As a rule of thumb, 70–80% compliance indicates good data quality. Answering almost all questionnaires (i.e., compliance close to 100%) can be a sign of reactivity, with participants adapting their daily life to make sure they do not miss a questionnaire. Adapting routines can compromise the ecological validity of an ESM study (Rintala et al., 2019). Nevertheless, whether high compliance compromises an ESM study may depend on the research aim and the length of the study (with longer durations having a higher risk for participants adapting to the questionnaire routine).

To guarantee a sufficient compliance and therefore a good data quality, several steps can be taken. First, by piloting the study, researchers can eradicate technical flaws and adapt the study design to prevent frustration (and potentially resulting in dropout) in their participants (Rintala et al., 2019; Eisele et al., 2020). Second, researchers can explain the importance of filling out most surveys to the participating parents before the study starts. Furthermore, presenting the participants with a manual of the device is associated with higher compliance rates (Morren et al., 2009). In our

experience, it is effective when researchers schedule an online or home visit with the participating parents to provide an explanation of the study application or device. Third, participants can be motivated to fill out many surveys with a (financial) reward system that can increase compliance rates (Morren et al., 2009). Note, however, to check within the sample what works as a reward. In our experience, mothers (mostly highly educated) were more motivated to help other parents (by contributing to science) or receive a present for their infant rather than to receive a monetary reward for themselves. In our study, we also added a “fun fact” about babies and/or parenting at the end of each questionnaire, which was indicated as one of the reasons that motivated mothers to continue completion of the daily questionnaires.

Most importantly, during data collection, researchers can ensure a high compliance by monitoring participation. Some software tools offer solutions to keep track of the surveys that are being filled out during the study. Seeing irregularities could be an indication of a technical problem. Therefore, it is advised to include the possibility to contact participants during the study to troubleshoot technical problems or answer questions of participants. Using text-based communications (i.e., purchasing a research-phone) can lower the threshold for participants to contact researchers in case of issues. In their review, Morren et al. (2009) also found that messages from the study researchers were an effective strategy to increase compliance.

Analyzing Your ESM Dataset

Experience sampling method enables the collection of intensive longitudinal data, with assessments nested within persons (and even within families). When analyzing ESM data, it is important to take into account these nested data structure in order to interpret the results correctly. Additionally, time plays an important role. The data are structured in a long format (vs. the wide data format) due to the multiple time points, with each row representing one time point per participant. Furthermore, for analyses, it is important to consider that assessments on day 1 are likely more strongly related to assessments on day 2 than on day 5 (van Roekel et al., 2019). Multilevel modeling is one analytic strategy that can be used to examine nested data and is often used in ESM studies. While specifying lagged variables is possible in multilevel models to take into account the time-dynamic structure, using software packages designed for examining these lagged dynamic associations is recommended (e.g., Dynamic Structural Equation Modeling (DSEM) in Mplus; (Asparouhov et al., 2018). DSEM is a statistical analysis technique that takes four methods of modeling into consideration that are suitable for ESM-specific data, namely, multilevel modeling, time-dynamic modeling, structural equation modeling, and time-varying effects modeling (Asparouhov et al., 2018). Therefore, DSEM can cater to ESM-specific data by catering to the unique aspects of such a study design and providing a complete picture of the study dynamics (Asparouhov et al., 2018). In addition, ESM studies

³<https://osf.io/2bm6x>

always have missing data, as 100% compliance is very unlikely (see section compliance). There are several methods to consider missing data in analyses. Therefore, it is important to consider missing data in the analyses of ESM data. For example, for multilevel analyses all available cases can be included, including participants who have not completed every point in time (Bagiella et al., 2000).

In the paper of Keijsers and Van Roekel (2018), suggestions of various methods for analyzing longitudinal data are presented as well as future recommendations, but the field is still under development. Most importantly is that the analyses should align with the research questions and that an adequate plan of analyses is made during the planning phase of the study. A full review of analyses techniques for ESM research is beyond the scope of the current article, and for more information, please see, among others, the studies by Asparouhov et al. (2018), Keijsers and van Roekel (2018), and Vogelsmeier et al. (2021).

COMMON PITFALLS AND THEIR SOLUTIONS

Like many other types of data collection, collecting data with ESM has its own complications. Because ESM heavily

relies on smartphones, many of the pitfalls have to do with or are at least somewhat related to the functioning of these devices or the functioning of the software. Even though the smartphones are quite literally out of our hands, there are still a number of ways in which researchers can avoid some common pitfalls with ESM research, both before and during the study. We discuss these in Box 1 and Box 2, respectively.

CONCLUSION

In sum, ESM is a relatively underused method for investigating infant development, even though it has multiple advantages over other research methods. Besides the possibility it offers to (micro)longitudinally study infant development from a distance, it also enables researchers to collect greatly detailed information about infant development and family dynamics in an ecologically valid manner. Highly dynamic concepts, such as mother-infant interactions, maternal mood, thoughts and behaviors, and infant sleep, cry, and fuss behaviors are particularly suitable for ESM. Nevertheless, transitioning into ESM research can be challenging and requires care, planning, and a commitment to the method – it cannot (and should not) merely be tagged onto existing research (Larson, 2019). With this practical guide, we hope to inform researchers involved in infant development about adding ESM to their research methods and to lower the threshold for incorporating it into their skillset.

AUTHOR CONTRIBUTIONS

MH coordinated the writing process, structured the outline, and wrote the introduction and conclusion. AB wrote paragraphs about ESM in general. VH wrote paragraphs about preregistration and participant recruitment and power analyses. LJ wrote paragraphs about wearables. MA wrote paragraphs about advantages and disadvantages. EM wrote paragraphs about common pitfalls and constructed the boxes. MB constructed the reference list. MH and MB monitored writing style, integrated all separate parts, and led the revisions. All authors provided feedback on the first complete draft and reviewed the final version of the manuscript.

FUNDING

This project was supported by awards from the Dutch Research Council (NWO), Veni.VI.191G.025 (PI: MH), a Vidi 452-17-011 (PI: Loes Keijsers), a Vici 453-15-006 (PI: Bernet Elzinga), and EM was supported by a grant from the European Research Council (ERC-2017-CoG-773023).

BOX 1 | Before the start of your study.

- ✓ Make sure that your study works as you want it to work. It is important to extensively test your study, by a number of people, using smartphones of different brands, makes, and ages.
- ✓ Consider the user experience. What is the length of your questionnaires? How often do you put out questionnaires a day? In short, what is the ESM burden?
 - Carefully consider if ESM burden is also dependent on the answers that your participants give (e.g., if they are alone vs. with others, do they not have to fill out any follow-up questions?). If this is the case, consider whether this is likely to influence the (truthfulness of) responses participants give.
- ✓ Check that people from your target population understand the questions and that they understand them in the way that you intended them.
- ✓ Consider your reward scheme; is the reward contingent on filling out a minimum number of questionnaires? Check whether such a scheme can also have negative consequences, for instance by demotivating participants who are not able to meet the minimum due to circumstances.
- ✓ Consider making a plan of analyses that are most suitable for the data that are going to be collected and to answer the research question appropriately. What are appropriate statistical models for these type of data?

BOX 2 | During the study.

- ✓ Make sure your participants know how to find you when something goes wrong. Consider a low-threshold medium for contact (e.g., SMS, WhatsApp, and call).
- ✓ Be prepared. Many ESM or daily diary applications will have a forum or help page. Make sure you are well-acquainted with it and already read up on some things that often go wrong.
- ✓ Keep track of compliance so that you can intervene when you see a participant is becoming less engaged. However, intervene with caution because too much interference can also decrease participants' motivation.

REFERENCES

- Adams, E. L., Marini, M. E., Brick, T. R., Paul, I. M., Birch, L. L., and Savage, J. S. (2019). Ecological momentary assessment of using food to soothe during infancy in the INSIGHT trial. *Int. J. Behav. Nutr. Phys. Act.* 16:79. doi: 10.1186/s12966-019-0837-y
- Ainsworth, J., Palmier-Claus, J. E., Machin, M., Barrowclough, C., Dunn, G., Rogers, A., et al. (2013). A comparison of two delivery modalities of a mobile phone-based assessment for serious mental illness: native smartphone application vs text-messaging only implementations. *J. Med. Internet Res.* 15:e60. doi: 10.2196/jmir.2328
- Asparouhov, T., Hamaker, E. L., and Muthén, B. (2018). Dynamic structural equation models. *Struct. Equ. Model. Multidiscip. J.* 25, 359–388. doi: 10.1080/10705511.2017.1406803
- Bacaro, V., Feige, B., Benz, F., Johann, A. F., De Bartolo, P., Devoto, A., et al. (2020). The association between diurnal sleep patterns and emotions in infants and toddlers attending nursery. *Brain Sci.* 10:891. doi: 10.3390/brainsci10110891
- Bagiella, E., Sloan, R. P., and Heitjan, D. F. (2000). Mixed-effects models in psychophysiology. *Psychophysiology* 37, 13–20. doi: 10.1111/1469-8986.3710013
- Bamberger, K. T. (2016). The application of intensive longitudinal methods to investigate change: stimulating the field of applied family research. *Clin. Child. Fam. Psychol. Rev.* 19, 21–38. doi: 10.1007/s10567-015-0194-6
- Barr, R. G., Kramer, M. S., Boisjoly, C., McVey-White, L., and Pless, I. B. (1988). Parental diary of infant cry and fuss behaviour. *Arch. Dis. Child.* 63, 380–387. doi: 10.1136/adc.63.4.380
- Boekhorst, M.G.B.M., Vergeer, J., Verhagen, C., van den Ende, D., and van den Heuvel, M. I. (2021). The Baby-related Anxiety and Behavior Inventory (BABI). 10.17605/OSF.IO/3TZNB.
- Bolger, N., Stadler, G., and Laurenceau, J.-P. (2011). “Power analysis for intensive longitudinal studies,” in *Handbook of Research Methods for Studying Daily Life*. M. R. Mehl and T. S. Conner (New York: Guilford), 285–301.
- Bolger, N., Davis, A., and Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annu. Rev. Psychol.* 54, 579–616. doi: 10.1146/annurev.psych.54.101601.145030
- Cho, E., and Ilari, B. S. (2021). Mothers as home DJs: recorded music and young children’s well-being during the COVID-19 pandemic. *Front. Psychol.* 12:637569. doi: 10.3389/fpsyg.2021.637569
- Christensen, T. C., Barrett, L. F., Bliss-Moreau, E., Lebo, K., and Kaschub, C. (2003). A practical guide to experience-sampling procedures. *J. Happiness Stud.* 4, 53–78. doi: 10.1023/A:1023609306024
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd Edn. (Hillsdale: Lawrence Erlbaum Associates).
- Consolvo, S., and Walker, M. (2003). Using the experience sampling method to evaluate ubicomp applications. *IEEE Pervasive Comput.* 2, 24–31. doi: 10.1109/MPRV.2003.1203750
- Corpus, R., D’Alessandro, S., and Collom, G. K. S. (2021). The postnatal testosterone rebound in first-time fathers and the quality and quantity of paternal care. *Dev. Psychobiol.* 63, 1415–1427. doi: 10.1002/dev.22064
- Csikszentmihalyi, M., and Larson, R. (1987). Validity and reliability of the experience-sampling method. *J. Nerv. Ment. Dis.* 175, 526–536. doi: 10.1097/00005053-198709000-00004
- de Jong, K., Moerbeek, M., and van der Leeden, R. (2010). A priori power analysis in longitudinal three-level multilevel models: an example with therapist effects. *Psychother. Res.* 20, 273–284. doi: 10.1080/10503300903376320
- Dietvorst, E., Hiemstra, M., Maciejewski, D., van Roekel, E., ter Bogt, T., Hillegers, M., et al. (2021). Grumpy or depressed? Disentangling typically developing adolescent mood from prodromal depression using experience sampling methods. *J. Adolesc.* 88, 25–35. doi: 10.1016/j.adolescence.2021.01.009
- Eisele, G., Vachon, H., Lافit, G., Kuppens, P., Houben, M., Myin-Germeyns, I., et al. (2020). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*. doi: 10.1177/1073191120957102 [Epub ahead of print]
- Fujiwara, T., Barr, R. G., Brant, R., and Barr, M. (2011). Infant distress at five weeks of age and caregiver frustration. *J. Pediatr.* 159, 425.e2–430.e2. doi: 10.1016/j.jpeds.2011.02.010
- Hajal, N. J., Teti, D. M., Cole, P. M., and Ram, N. (2019). Maternal emotion, motivation, and regulation during real-world parenting challenges. *J. Fam. Psychol.* 33, 109–120. doi: 10.1037/fam0000475
- Hechler, C., Beijers, R., Riksen-Walraven, J. M., and de Weerth, C. (2018). Are cortisol concentrations in human breast milk associated with infant crying? *Dev. Psychobiol.* 60, 639–650. doi: 10.1002/dev.21761
- Hektner, J. M., and Csikszentmihalyi, M. (2002). “The experience sampling method: measuring the context and the content of lives,” in *Handbook of Environmental Psychology* (Hoboken: John Wiley & Sons, Inc.), 233–243.
- Himmelstein, P. H., Woods, W. C., and Wright, A. G. C. (2019). A comparison of signal- and event-contingent ambulatory assessment of interpersonal behavior and affect in social situations. *Psychol. Assess.* 31, 952–960. doi: 10.1037/pas0000718
- Kaufmann, K., and Peil, C. (2019). The mobile instant messaging interview (MIMI): using WhatsApp to enhance self-reporting and explore media usage in situ. *Mobile Media Commun.* 8, 229–246. doi: 10.1177/2050157919852392
- Keijsers, L., and van Roekel, E. (2018). “Longitudinal methods in adolescent psychology: where could we go from here? And should we?” in *Reframing Adolescent Research*. 1st Edn. eds. L. B. Hendry and M. Kloep (London: Routledge).
- Kim, S., Holloway, S. D., Bempechat, J., and Li, J. (2018). Explaining adolescents’ affect: A time-use study of opportunities for support and autonomy across interpersonal contexts. *J. Child Fam. Stud.* 27, 2384–2393. doi: 10.1007/s10826-018-1092-6
- Kirtley, O. J., Lافit, G., Achterhof, R., Hiekkaranta, A. P., and Myin-Germeyns, I. (2021). Making the black box transparent: A template and tutorial for registration of studies using experience-sampling methods. *Adv. Methods Pract. Psychol. Sci.* 4:251524592092468. doi: 10.1177/2515245920924686
- Lافit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeyns, I., Viechtbauer, W., and Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Adv. Methods Pract. Psychol. Sci.* 4:251524592097873. doi: 10.1177/2515245920978738
- Larson, R. W. (2019). Experiencing sampling research from its beginnings into the future. *J. Res. Adolesc.* 29, 551–559. doi: 10.1111/jora.12524
- Larson, R. W., Richards, M. H., Moneta, G., Holmbeck, G., and Duckett, E. (1996). Changes in adolescents’ daily interactions with their families from ages 10 to 18: disengagement and transformation. *Dev. Psychol.* 32, 744–754. doi: 10.1037/0012-1649.32.4.744
- Mindell, J. A., and Lee, C. (2015). Sleep, mood, and development in infants. *Infant Behav. Dev.* 41, 102–107. doi: 10.1016/j.infbeh.2015.08.004
- Morren, M., van Dulmen, S., Ouwkerk, J., and Bensing, J. (2009). Compliance with momentary pain measurement using electronic diaries: a systematic review. *Eur. J. Pain* 13, 354–365. doi: 10.1016/j.ejpain.2008.05.010
- Myin-Germeyns, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., et al. (2018). Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry* 17, 123–132. doi: 10.1002/wps.20513
- Naab, T. K., Karnowski, V., and Schlütz, D. (2019). Reporting Mobile social media use: how survey and experience sampling measures differ. *Commun. Methods Meas.* 13, 126–147. doi: 10.1080/19312458.2018.1555799
- Palmier-Claus, J. E., Myin-Germeyns, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P. A., et al. (2011). Experience sampling research in individuals with mental illness: reflections and guidance. *Acta Psychiatr. Scand.* 123, 12–20. doi: 10.1111/j.1600-0447.2010.01596.x
- Repetti, R. L., Reynolds, B. M., and Sears, M. S. (2015). Families under the microscope: repeated sampling of perceptions, experiences, biology, and behavior. *J. Marriage Fam.* 77, 126–146. doi: 10.1111/jomf.12143
- Rintala, A., Wampers, M., Myin-Germeyns, I., and Viechtbauer, W. (2019). Response compliance and predictors thereof in studies using the experience sampling method. *Psychol. Assess.* 31, 226–235. doi: 10.1037/pas0000662
- Salo, V. C., Pannuto, P., Hedgecock, W., Biri, A., Russo, D. A., Piersiak, H. A., et al. (2021). Measuring naturalistic proximity as a window into caregiver-child interaction patterns. *Behav. Res. Methods*. doi: 10.3758/s13428-021-01681-8 [Epub ahead of print]

- Sanjuan, P. M., Pearson, M. R., Fokas, K., and Leeman, L. M. (2020). A mother's bond: An ecological momentary assessment study of posttraumatic stress disorder symptoms and substance craving during pregnancy. *Psychol. Addict. Behav.* 34, 269–280. doi: 10.1037/adb0000543
- Sawada, N., Gagné, F. M., Séguin, L., Kramer, M. S., McNamara, H., Platt, R. W., et al. (2015). Maternal prenatal felt security and infant health at birth interact to predict infant fussing and crying at 12 months postpartum. *Health Psychol.* 34, 811–819. doi: 10.1037/hea0000152
- Schwarz, N. (2007). "Retrospective and concurrent self-reports: the rationale for real-time data capture," in *The Science of Real-Time Data Capture: Self-Reports in Health Research*. eds. A. Stone, S. Shiffman, A. Atienza and L. Nebeling (New York: Oxford University Press), 11–26.
- Scollon, C. N., Kim-Prieto, C., and Diener, E. (2003). Experience sampling: promises and pitfalls, strengths and weaknesses. *J. Happiness Stud.* 4, 5–34. doi: 10.1023/A:1023605205115
- Scollon, C., Prieto, C., and Diener, E. (2009). "Experience sampling: promises and pitfalls, strength and weaknesses," in *Assessing Well-Being. Social Indicators Research Series (Vol. 39)*. ed. E. Diener (Dordrecht: Springer).
- Sonnenberg, B., Riediger, M., Wrzus, C., and Wagner, G. G. (2012). Measuring time use in surveys – concordance of survey and experience sampling measures. *Soc. Sci. Res.* 41, 1037–1052. doi: 10.1016/j.ssresearch.2012.03.013
- St James-Roberts, I., and Plewis, I. (1996). Individual differences, daily fluctuations, and developmental changes in amounts of infant waking, fussing, crying, feeding, and sleeping. *Child Dev.* 67, 2527–2540. doi: 10.2307/1131638
- Stieger, S., and Reips, U.-D. (2019). Well-being, smartphone sensors, and data from open-access databases: A mobile experience sampling study. *Field Methods* 31, 277–291. doi: 10.1177/1525822X18824281
- Stone, A. A., and Shiffman, S. (1994). Ecological momentary assessment (EMA) in behavioral medicine. *Ann. Behav. Med.* 16, 199–202. doi: 10.1093/abm/16.3.199
- Teti, D. M., Kim, B.-R., Mayer, G., and Counterline, M. (2010). Maternal emotional availability at bedtime predicts infant sleep quality. *J. Fam. Psychol.* 24, 307–315. doi: 10.1037/a0019306
- Thomas, V., and Azmitia, M. (2015). Tapping Into the app: updating the experience sampling method for the 21st century. *Emerg. Adulthood* 4, 60–67. doi: 10.1177/2167696815618489
- Tikotzky, L., and Sadeh, A. (2009). Maternal sleep-related cognitions and infant sleep: a longitudinal study from pregnancy through the 1st year. *Child Dev.* 80, 860–874. doi: 10.1111/j.1467-8624.2009.01302.x
- Trull, T. J., and Ebner-Priemer, U. W. (2009). Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. *Psychol. Assess.* 21, 457–462. doi: 10.1037/a0017653
- Uy, M. A., Foo, M.-D., and Aguinis, H. (2009). Using experience sampling methodology to advance entrepreneurship theory and research. *Organ. Res. Methods* 13, 31–54. doi: 10.1177/1094428109334977
- van Berckel, N., Ferreira, D., and Kostakos, V. (2017). The experience sampling method on Mobile devices. *ACM Comput. Surv.* 50, 1–40. doi: 10.1145/3123988
- van Roekel, E., Keijsers, L., and Chung, J. M. (2019). A review of current ambulatory assessment studies in adolescent samples and practical recommendations. *J. Res. Adolesc.* 29, 560–577. doi: 10.1111/jora.12471
- Vanden Abeele, M. M. P., Abels, M., and Hendrickson, A. T. (2020). Are parents less responsive to young children when they are on their phones? A systematic naturalistic observation study. *Cyberpsychol. Behav. Social Networking* 23, 363–370. doi: 10.1089/cyber.2019.0472
- Vogelsmeier, L., Vermunt, J. K., Keijsers, L., and De Roover, K. (2021). Latent Markov latent trait analysis for exploring measurement model changes in intensive longitudinal data. *Eval. Health Prof.* 44, 61–76. doi: 10.1177/0163278720976762
- Wen, C. K. F., Schneider, S., Stone, A. A., and Spruijt-Metz, D. (2017). Compliance with mobile ecological momentary assessment protocols in children and adolescents: A systematic review and meta-analysis. *J. Med. Internet Res.* 19:e132. doi: 10.2196/jmir.6641
- Williams, M. T., Lewthwaite, H., Fraysse, F., Gajewska, A., Ignatavicius, J., and Ferrar, K. (2021). Compliance with mobile ecological momentary assessment of self-reported health-related behaviors and psychological constructs in adults: systematic review and meta-analysis. *J. Med. Internet Res.* 23:e17023. doi: 10.2196/17023

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 van den Heuvel, Bülow, Heininga, de Moor, Janssen, Vanden Abeele and Boekhorst. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Bringing Home Baby Euclid: Testing Infants' Basic Shape Discrimination Online

Agata Bochynska and Moira R. Dillon*

Department of Psychology, New York University, New York City, NY, United States

OPEN ACCESS

Edited by:

Rhodri Cusack,
Trinity College Institute of
Neuroscience, Ireland

Reviewed by:

Paul Muentener,
Tufts University, United States
J. Kiley Hamlin,
University of British Columbia,
Canada
Caroline Junge,
Utrecht University, Netherlands

*Correspondence:

Moira R. Dillon
moira.dillon@nyu.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 01 July 2021

Accepted: 25 November 2021

Published: 20 December 2021

Citation:

Bochynska A and Dillon MR (2021)
Bringing Home Baby Euclid: Testing
Infants' Basic Shape Discrimination
Online.
Front. Psychol. 12:734592.
doi: 10.3389/fpsyg.2021.734592

Online developmental psychology studies are still in their infancy, but their role is newly urgent in the light of the COVID-19 pandemic and the suspension of in-person research. Are online studies with infants a suitable stand-in for laboratory-based studies? Across two unmonitored online experiments using a change-detection looking-time paradigm with 96 7-month-old infants, we found that infants did not exhibit measurable sensitivities to the basic shape information that distinguishes between 2D geometric forms, as had been observed in previous laboratory experiments. Moreover, while infants were distracted in our online experiments, such distraction was nevertheless not a reliable predictor of their ability to discriminate shape information. Our findings suggest that the change-detection paradigm may not elicit infants' shape discrimination abilities when stimuli are presented on small, personal computer screens because infants may not perceive two discrete events with only one event displaying uniquely changing information that draws their attention. Some developmental paradigms used with infants, even those that seem well-suited to the constraints and goals of online data collection, may thus not yield results consistent with the laboratory results that rely on highly controlled settings and specialized equipment, such as large screens. As developmental researchers continue to adapt laboratory-based methods to online contexts, testing those methods online is a necessary first step in creating robust tools and expanding the space of inquiry for developmental science conducted online.

Keywords: change detection, geometry, online study, shape perception, infants

INTRODUCTION

Online studies with adults have been around in psychological research for many years, and many web-based solutions have been validated for adult testing (Buhrmester et al., 2011; Crump et al., 2013; de Leeuw et al., 2014; Gureckis et al., 2016; Sauter et al., 2020). Online studies with infants and children, however, are a relatively recent development that became newly urgent in the light of the COVID-19 pandemic and the suspension of in-person research (Lourenco and Tasimi, 2020; Sheskin et al., 2020; Zaadnoordijk et al., 2021). Because infants and young children cannot simply read the instructions and click through web-based tasks unsupervised, different solutions have been proposed for collecting developmental data online. For example, commercial or custom-built video-chat software allows an experimenter to interact

with a participant through a webcam in real time while running a study remotely (Sheskin and Keil, 2018). Online platforms for unmoderated developmental research (Scott et al., 2017; Scott and Schulz, 2017; Rhodes et al., 2020; Lo et al., 2021); moreover, present detailed instructions addressed to parents or guardians allowing them to participate with their children from their home computer with a webcam, without the experimenter present and without an appointment. Several questions naturally arise: Is there a difference between online and in-laboratory results? Are there comparative advantages or unique limitations to either context? Indeed, can we ask new questions now that the space of inquiry has expanded?

Several recent online studies have found results that are mostly consistent with in-laboratory results using either the moderated video-chat or the unmoderated approach. These studies have nevertheless adapted forced-choice paradigms with children or looking-time paradigms with older infants and toddlers like preferential or “violation-of-expectation” paradigms (Scott et al., 2017; Sheskin and Keil, 2018; Leshin et al., 2020; Nussenbaum et al., 2020; Lo et al., 2021; Smith-Flores et al., 2021). Such results thus do not address whether other common methods used in developmental research, for example, some looking-time paradigms with younger infants, may be adaptable to online contexts and serve as a replacement for in-person, laboratory testing. In the present study, we thus ask whether certain early emerging abilities to discriminate shape information, which is foundational both for infants’ everyday interactions with objects (e.g., Quinn and Eimas, 1997; Quinn et al., 2001; Smith, 2009) as well as for children’s later achievement in STEM (Science, Technology, Engineering, and Mathematics) fields (e.g., Verdine et al., 2017), and which have been revealed through highly controlled laboratory studies with specialized setups and equipment, might also be measurable using unmonitored online testing, relying only on a personal computer with a webcam.

To address this question, we adapted two experiments with 7-month-old infants from a series of experiments on infant shape discrimination conducted in a laboratory setting (Dillon et al., 2020). These experiments used a “change-detection” looking-time paradigm (after Ross-Sheehy et al., 2003), in which rapidly changing displays presented visual forms (triangles or open “V” figures) on a large projector screen. On one side of the screen, the visual forms were changing in shape and area, while on the other side of the screen, the visual forms were changing in area only. On both sides, the figures were additionally changing in position and orientation. The rationale behind this paradigm is that if infants look longer at the stream of figures with the one additional change (in this case, the shape change), then that serves as evidence of their detection of that change. Dillon et al. (2020) observed across four experiments that infants showed significantly more looking to the figure streams with a shape-and-area change compared to an area-only change in full triangles and in “V” figures with relative length changes.

The change-detection paradigm has been used to investigate a variety of infants’ abilities in laboratory settings, including their sensitivity to mirror reversals in visual forms (Lauer et al., 2015), to numerical differences in dot arrays (Libertus

and Brannon, 2010; Schröder et al., 2020), and to bound color and object information in visual short-term memory (Ross-Sheehy et al., 2003). It has also been used in the laboratory to chart developmental changes in infancy, e.g., in numerical discrimination from 6 to 9 months (Libertus and Brannon, 2010) and in visual short-term memory from 4 to 13 months (Ross-Sheehy et al., 2003). Moreover, small-scale longitudinal studies in the laboratory have relied on change detection to measure infants’ individual sensitivities to number and geometry, and these studies have revealed stable change detection across individuals in infancy and correlations between change detection in infancy and performance on standardized measures of symbolic mathematics in young childhood (Starr et al., 2013; Lauer and Lourenco, 2016). With the possibility that online testing will allow for larger sample sizes and the ability to collect repeated measures with the same infants over longer periods of time compared to in-laboratory testing (Sheskin et al., 2020), change detection thus becomes a prime candidate for supporting large-scale, longitudinal studies focusing on development and individual differences across domains.

The change-detection paradigm, moreover, offers additional scientific and practical advantages relative to other looking-time paradigms used with infants, like the “habituation” paradigm, which has also been used extensively in the laboratory to measure infants’ numerical and spatial sensitivities. For example, studies using habituation to evaluate infants’ shape discrimination (Schwartz and Day, 1979; Cohen and Younger, 1984; Slater et al., 1991) have relied on long presentations times, considerably longer than those reflected in natural viewing (Yu and Smith, 2016). The rapid displays used in change detection, in contrast, better reflect the dynamically changing visual world of infants’ everyday life. Moreover, the change-detection paradigm may result in lower numbers of excluded participants (Lauer et al., 2015), permits the use of other measurement tools such as automated eye tracking, and relies on fixed-duration presentations, which allow for offline coding, fewer research personnel, and even remote, unmonitored data collection.

It nevertheless remains an open question whether change detection can be adopted for online testing. In particular, most studies using change detection in the laboratory have relied on stimuli being presented on two separate monitors (e.g., Ross-Sheehy et al., 2003; Libertus and Brannon, 2010) or on a very large projector screen (e.g., Lauer et al., 2015; Dillon et al., 2020), neither of which are typically present in the home. Those that have relied on smaller screens (e.g., Schröder et al., 2020) have failed to find some of the same change-detection capacities that were found with larger screens, and unpublished data suggest that change-detection findings in the numerical domain measured in the laboratory may not robustly replicate, on either small or large screens (Lindskog et al., unpublished data). In the present study, we thus ask whether robust in-laboratory findings using change detection that presented rapidly changing, simple 2D figures on a large screen could be found using unmonitored, online data collection with stimuli presented at home on small, personal computer screens.

The present study includes two sequential experiments, one modeled after Experiment 1B and one modeled after 2B from

Dillon et al. (2020). Both of these experiments produced robust findings in the laboratory; they were replications and extensions of Experiment 1A and Experiment 2A also from Dillon et al. (2020). All four of these experiments, moreover, yielded similar and medium-to-large effect sizes (Cohen's *ds*: Experiment 1A: 0.71; Experiment 1B: 0.66; Experiment 2A: 0.68; Experiment 2B: 0.98). The methods and analysis plans for both of the present experiments were preregistered on the Open Science Framework prior to data collection,¹ and the data and analysis code are publicly available (data: <https://osf.io/ecyfd/>; analysis code: <https://osf.io/munk7/>). The first experiment was conducted on the unmonitored online developmental testing platform Lookit² when Lookit was still under development and was accessible only to a limited number of researchers. The second experiment was also conducted on Lookit, but after its beta testing had been completed and during its transition to a platform accessible to those able to comply with Lookit's access agreement.

GENERAL METHODS

Families participated in one of two experiments through the online developmental testing platform Lookit (Scott and Schulz, 2017). They were mainly recruited by phone or email from databases of families who had expressed interest in participating in research studies, one database at Harvard University and two databases at New York University. Families were also recruited from Lookit's participant database, posted flyers, online forums, social media sites, and word-of-mouth. They received a \$5 Amazon gift card for participating. Our use of human participants was approved by the Institutional Review Boards at Massachusetts Institute of Technology (MIT; cede agreement for multi-site research at MIT and Harvard University) and at New York University. Our use of Lookit was approved initially under this cede agreement and then under Lookit's access agreement.

The materials and design of the experiments are illustrated in **Figure 1**. After Dillon et al. (2020), the experiments followed a change-detection paradigm in which dynamic streams of 2D figures appeared simultaneously, one on the left side of the screen and one on the right side of the screen, each stream bounded by a static rectangle. One stream presented figures changing in area alone (a shape-preserving scale transformation), and the other stream presented figures changing in shape and area (a shape change that resulted in an area change). For half of the infants, the area change resulted in smaller figures, and for the other half of the infants, the area change resulted in larger figures. On both sides of the screen, there was additional random variation in the figures' positions ($\pm 4.5\%$ relative to the center of the bounding rectangle in both the vertical and horizontal directions), orientations ($\pm 0-359^\circ$), and sizes (a shape-preserving scale transformation $\pm 15\%$). While we had planned to vary figures' left-right direction

randomly for each presentation in both experiments, this variation was only implemented in Experiment 2 because of an error in Lookit's stimuli presentation code. Each figure in each stream appeared for 0.5 s followed by a 0.3-s blank screen before the next figure appeared. Streams were presented in four 60-s blocks, and the shape change appeared on alternating sides of the screen across blocks. The shape change started on the left side of the screen for half of the infants, and it started on the right side of the screen for the other half of the infants.

Dillon et al. (2020) presented forms as light blue outlines on a black background projected on a large screen (1.07 m \times 1.37 m) in a dimly lit quiet laboratory testing room. Parents sat 1.70 m from the screen, positioned infants on their laps, and closed their eyes during the stimuli presentation. They received live instruction from an experimenter who stood behind the screen and came out after each trial to reset the infant on the parent's lap – if needed – and to recalibrate the infant's looking. During calibration, the experimenter shook a rattle in front of different locations on the screen. Before the stimuli started, a pink circle appeared in the center of the screen, and the experimenter used the infant's name to draw their attention to the circle (see <https://osf.io/b3g52/> for example stimuli). The test trials were silent.

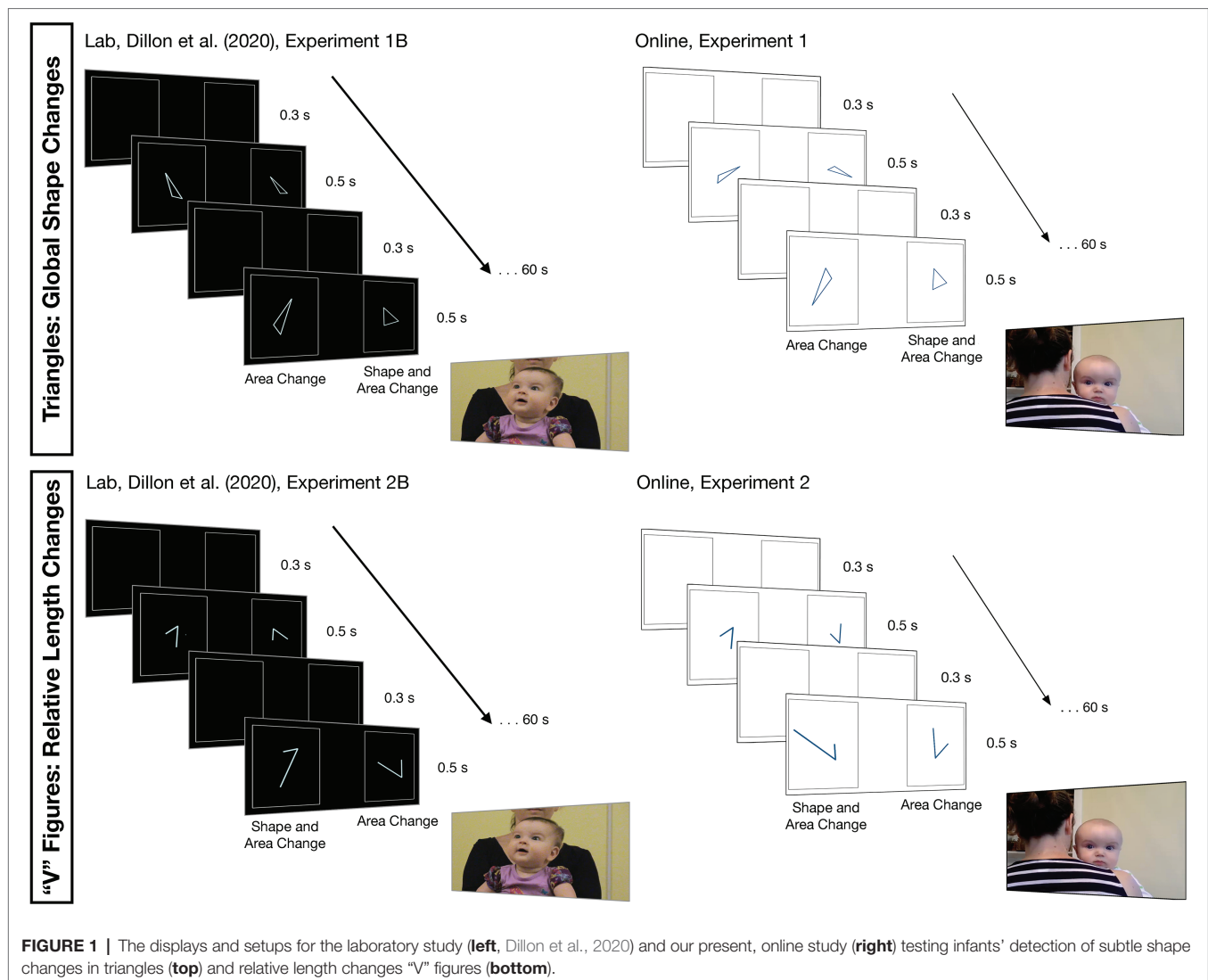
The differences between Dillon et al. (2020) and the present experiments are illustrated in **Figure 1**. In contrast to Dillon et al. (2020), our stimuli flexibly scaled to fit the screen of the personal computer on which they were being presented. To maximize visibility in the variable lighting conditions of the home-testing environment, moreover, we presented forms as dark blue outlines on a white background. Parents sat about an arm-length distance from the screen, faced away from the screen, and held their infants over their shoulders to face the screen. Our experiment, moreover, was completely unmonitored. Parents followed a set of written and pictorial descriptions instructing them how to set up the home-testing environment. Pre-recorded audio specifying the start of the experiment, the trial number, and the end of the experiment guided parents, and a twirling star with an accompanying chime sound appeared at different locations on the screen prior to each trial to calibrate infants' looking and to introduce the test trials. Test trials were accompanied by looping infant-friendly music.

GENERAL PREREGISTERED ANALYSIS

Coding and analyses followed Dillon et al. (2020). In both experiments, we measured infants' total looking time to the figure stream presenting changes in shape and area and the stream presenting changes in area alone. Infants' looking time to the streams was coded offline in real time from digital video recordings by a researcher masked to the changes that the infant was seeing. The total looking of 12 random infants in each experiment (25%) was recoded in their same way by a different researcher. For each infant, we calculated their proportion of looking to the shape-and-area-change stream as a function of their total looking to both streams across all

¹<https://osf.io/vvaw7/registrations>

²<https://lookit.mit.edu/>



four trials (see also Lauer et al., 2015). This proportion was compared to 0.50 using a one-sample, two-tailed *t*-test.

EXPERIMENT 1

Experiment 1 adapted Experiment 1B from Dillon et al. (2020) and explored whether infants detected global shape changes in closed 2D triangles, over and above changes in triangle position, orientation, and size.

Methods

Participants

Data collection took place from late March 2017 to early April 2018. Forty-eight full-term (≥ 37 weeks gestational age at birth) 7-month-old infants were included in the sample (22 females, mean age = 7 months 3 days, range = 6 months 15 days to 7 months 15 days). The planned sample size of 48 infants was preregistered based on a power analysis of the findings of Dillon et al. (2020),

Experiment 1B (with Cohen's $d=0.66$ and $SD=0.07$); power was 99.4%. For 51 families who completed the consent video, we received no test videos, and for four additional families, we received only partial test videos. Informal parental reports and discussions with the Lookit staff suggested that technical difficulties led to this large amount of missing data (due, in particular, to Lookit's running on Adobe Flash, not HTML5, at the time). Five families completed at least one but fewer than four test blocks, and one family had poor video quality. Two additional families withdrew their consent before participating. In the corresponding laboratory study from Dillon et al., 2020, which had a sample size of 16 infants, no additional infants were excluded.

Displays

After Experiment 1B of Dillon et al., 2020, four triangles were used as stimuli: two similar 45° - 60° - 75° triangles and two similar 15° - 45° - 120° triangles. The areas of the smaller and larger versions of each triangle were matched across the two triangle types and differed by a factor of two

(see Dillon et al., 2020, for additional details on the geometric properties of the stimuli). Each infant saw three of the four triangles. Half of the infants saw the larger 45°-60°-75° triangle on both sides of the screen, alternating with the smaller 15°-45°-120° in the shape-and-area-change stream and the smaller 45°-60°-75° triangle in the area-only-change stream. The other half of the infants saw the smaller 15°-45°-120° on both sides of the screen, alternating with the larger 45°-60°-75° triangle in the shape-and-area-change stream and the larger 15°-45°-120° triangle in the area-only-change stream.

Results

Primary Preregistered Analysis

We preregistered the specification that if parents watched the test stimuli for 1 s or more, infants' looking times would be included only up until the point at which their parent watched the stimuli for that particular block. Before analyzing our data, however, we decided to include infants' looking times even if their parent watched the test stimuli. This more inclusive analysis, which we report in the main text, is consistent with the planned analysis, and so we report the planned analysis in the **Supplementary Material**.

The reliability of the two looking-time coders was high (Pearson $r=0.94$). Unlike in Dillon et al. (2020) Experiment 1B, infants did not look significantly longer to the shape-and-area-change stream compared to the area-only-change stream [$t(47)=0.27$, $p=0.786$, $d=0.04$; **Figure 2A**].

Secondary Preregistered Analysis

To further understand our findings, we first identified any influential participants by calculating Cook's distance in a linear regression on raw total looking times to each stream for each infant with Change Type (shape-and-area change or area-only change) as a fixed effect. The analysis identified two influential participants. We reran the primary analysis after excluding these participants, and our results were consistent with the primary analysis [$t(45)=0.41$, $p=0.683$, $d=0.06$].

Next, we ran a mixed-model linear regression on infants' raw looking times after the model from Dillon et al. (2020). We had misspecified this model in our preregistration, and the correct model included Change Type (shape-and-area change or area-only change), Size (bigger triangle or smaller triangle), Block (1, 2, 3, or 4), and Gender as fixed effects, and Participant as a random-effects intercept. Consistent with the primary analysis, we found no significant effect of Change Type ($\beta=0.65$, $p=0.503$), again providing no evidence that infants looked longer to the shape-and-area-change stream compared to the area-only-change stream. There were also no significant effects of Size ($\beta=1.21$, $p=0.363$) or Gender ($\beta=1.02$, $p=0.443$), and consistent with Dillon et al. (2020), there was a significant effect of Block ($\beta=-2.22$, $p<0.001$), with looking time decreasing across blocks. An additional regression using this model with incomplete datasets (we received five such datasets, but three had a condition assignment that we could not determine) showed results consistent with the primary analysis and so are reported in the **Supplementary Material**.

Finally, to examine whether any effects might be measurable from experiments that are shorter in duration (and thus perhaps more adaptable to online sessions) we repeated our primary analysis but only considered the first two blocks. This analysis also showed that infants did not look significantly longer to the shape-and-area-change stream compared to the area-only-change stream [$t(47)=0.69$, $p=0.495$, $d=0.10$].

Exploratory Analysis

Our exploratory analysis specifically aimed to examine the differences between the present results and the results in Dillon et al. (2020), Experiment 1B. A direct comparison between the two experiments using an independent samples t -test found a significant difference between infants' preference for the shape-and-area-change stream across the two experiments [$t(27)=2.10$, $p=0.045$, $d=0.59$]. Given that our experiment differed from the original experiment in many respects as outlined above, our exploratory analyses thus focus on evaluating any effects of those differences, where possible.

First, infants looked longer at the stimuli online compared to the laboratory [$t(33.57)=-4.78$, $p<0.001$, $d=1.21$], suggesting that infants at least saw the stimuli for a long enough time to show the expected effect. Second, parents tested online were instructed to hold their infants over their shoulders as opposed to on their laps, and this position may have resulted in longer looking to the side of the screen away from the parent's head, biasing the overall pattern of results. That said, about half (26/48) of parents held their child over their left shoulder for the duration of the study and three parents switched sides, so, across infants, neither side of the screen was potentially more visually accessible. Accordingly, a mixed model linear regression with Change Type (shape-and-area change or area-only change) and Side (left or right) revealed no significant effect of Side ($\beta=-0.06$, $p=0.966$) and no Change Type X Side interaction ($\beta=-1.22$, $p=0.542$).

Next, we focused on exploring infants' distraction, which may have uniquely affected their ability to detect shape changes in an uncontrolled at-home environment versus a highly controlled laboratory environment. Following Scott and Schulz (2017), a researcher, masked to what infants saw and their individual looking times, recoded the videos to enumerate the following types of distracting events: fussiness (e.g., crying or squirming to get out of a parent's lap); distracted by an external event (e.g., someone walking by); and distracted by an external object (e.g., dropping a toy or pacifier; see Scott and Schulz, 2017, for additional details). Twenty-three of the 48 infants in our sample experienced at least one distracting event ($M=3.61$; $Median=2$) during the experiment. A Spearman correlation revealed that the number of distracting events negatively correlated with infants' overall looking time ($r_s=-0.57$, $p<0.001$). Surprisingly, a Spearman correlation also revealed that the number of distracting events negatively correlated with the proportion looking to the shape-and-area-change stream across infants ($r_s=-0.32$, $p=0.025$). Infants who had one or fewer distracting events ($N=34$), moreover, showed a positive, although not significant, preference for the shape-and-area-change stream

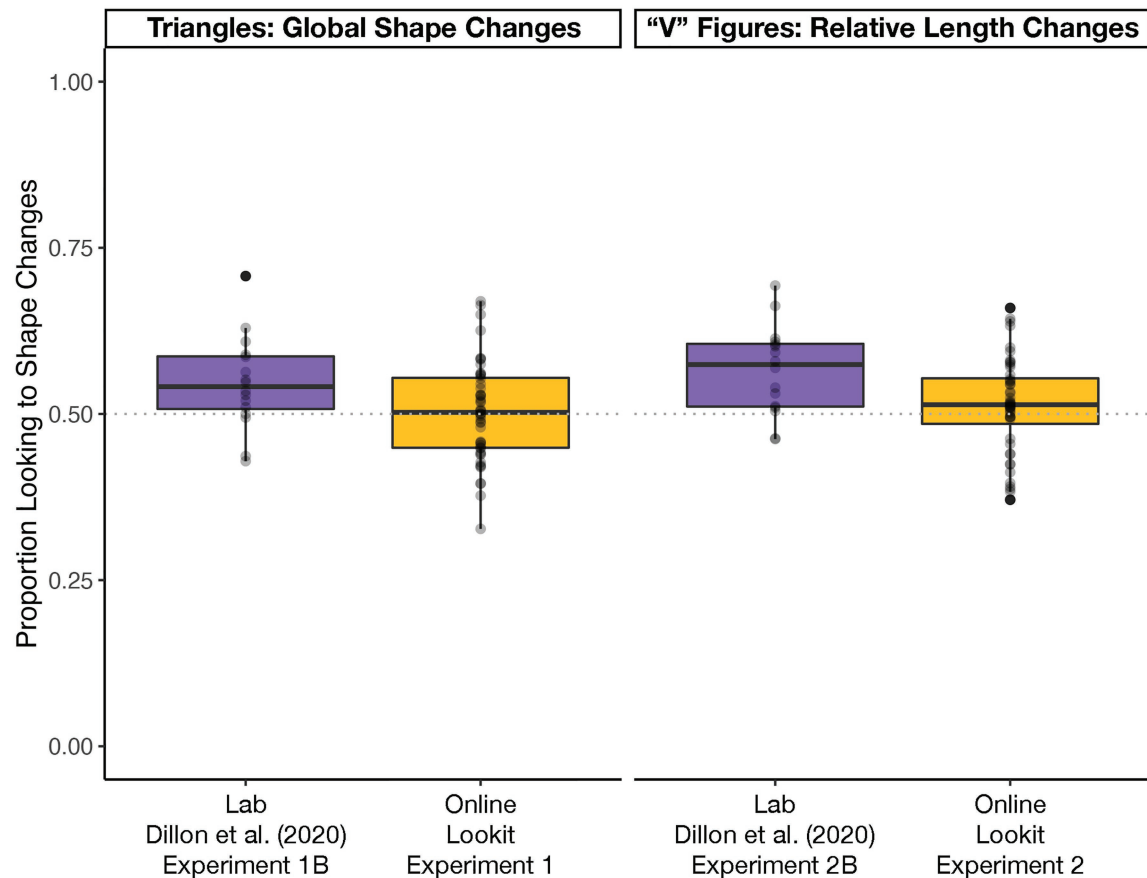


FIGURE 2 | Boxplots describing the proportions of infants' looking to shape changes (**left**) and relative length changes (**right**) in laboratory experiments, reported in Dillon et al. (2020, $N=16$ per experiment), and in the present online experiments ($N=48$ per experiment). They gray-dotted line at 0.50 indicates no looking preference, and the overlaid points display each participant's individual preference, collapsed across an experiment's four blocks. While infants looked longer at shape changes in the laboratory experiments, they did not look longer at shape changes in the online experiments.

[$M=0.52$, $SEM=0.01$, $t(33)=1.73$, $p=0.093$, $d=0.30$]. These results suggest that distraction might explain why infants in the online experiment did not show the same patterns of change detection of visual forms as were observed in the laboratory studies.

Discussion

Unlike in Dillon et al. (2020), in Experiment 1, we found no evidence that 7-month-old infants looked significantly longer to shape changes in triangles over and above changes in triangle position, orientation, and size. To further understand how the results from Dillon et al.'s (2020) laboratory study compared to our present online study, we explored the possibility that infants were distracted in the home environments and that this distraction affected infants' ability to detect subtle shape changes in rapidly presented displays of 2D figures. We found that the number of times that infants were distracted during the stimuli presentation negatively correlated with their ability to detect shape changes.

In Experiment 2, we thus focused on two aims. First, we focused on an experiment from Dillon et al. (2020) that

probed infants' detection of relative length changes instead of global shape changes. Relative length changes are also robustly detected in the laboratory, they can be instantiated in simpler 2D figures, and they may underlie infants' detection of global shape changes in triangles (Dillon et al., 2020). Second, inspired by the exploratory analysis of Experiment 1, we focused on distraction specifically as the cause of the difference between the findings of the in-laboratory versus online versions of the experiment. We did so by improving the instructions given to the parents to minimize possible distractions in the home, and we introduced new exclusion criterion based on distraction, with planned analyses that allowed us evaluate the effects of distraction directly.

EXPERIMENT 2

Experiment 2 adapted Experiment 2B from Dillon et al. (2020) and explored whether infants detected shape changes, instantiated as changes to the relative lengths of the arms forming open 2D "V" figures, over and above changes in figure position, orientation, sense, and size. As a result of the methods and exploratory findings

of Experiment 1, Experiment 2 also included improved instructions to parents and preregistered exclusion criteria and analyses based on infant distraction and potential parent interference.

Methods

Participants

Data collection took place from late March 2020 to late November 2020. Because of the null result in Experiment 1 and how resource intensive data collection is with infants, we preregistered a sequential sampling procedure. We used a Bayes Factor Design Analysis for sequential designs (see Stefan et al., 2019) using $r = \sqrt{2}/2$ as a default prior distribution on effect size δ , which we estimated as 0.35. After every eight infants who met the inclusion criteria, we evaluated a one-sample Bayesian t -test with a directional hypothesis, comparing infants' proportion of looking to the shape-and-area-change stream to 0.50. We aimed for a strength of evidence of 6, which meant that we would collect data until the Bayes Factor was larger than 6 (evidence for H_1), smaller than $1/6$ (evidence for H_0), or we reached a maximum sample size of 48 infants.

We did not meet the planned strength of evidence before reaching the maximum sample size, and so data from 48 full term (≥ 37 weeks gestational age at birth) 7-month-old infants were included in the sample (24 females, mean age = 6 months 28 days, range = 6 months 15 days to 7 months 15 days).

For 12 families who completed the consent video, we received no test videos, and for 34 additional families, we received only partial test videos. Discussions with the Lookit staff suggested that technical difficulties led to this large amount of missing data. Three families completed at least one but fewer than four blocks, and three families had poor video quality. In addition, a large number of infants (42) were excluded based on the new preregistered exclusion criteria motivated by the exploratory analysis of the effects of infant distraction from Experiment 1: eight infants with looking times < 80 s; seven infants with parents who watched the test stimuli; nine infants who were distracted; two infants with looking times < 80 s who were distracted; one infant with looking time < 80 s whose parents watched the test stimuli; two infants with looking times < 80 s who were distracted and whose parents watched the test stimuli; and 11 infants who were distracted and had parents who watched the test stimuli. In the corresponding laboratory study from Dillon et al., 2020 ($N = 16$) one additional infant was excluded because of low looking time, one because of a preference score of more than two standard deviations above or below the mean, and one because of equipment failure.

Displays

After Dillon et al., 2020, Experiment 2B, four "V" figures were used, all with an angle measure of 53.39° (see Figure 1). Two of those figures had an arm-length ratio of 1:1.5 and two had an arm-length ratio of 1:3, so the relative length difference between the two figure types was 1:2. For each of the two figure types, there was one version that has a smaller implied area (formed by joining the endpoints at the open side of the "V" to make a triangle) and one that has a larger implied area (see

Dillon et al., 2020, for additional details on the geometric properties of the stimuli). Each infant saw three of the four "V" figures. Half of the infants saw the larger 1:3 "V" figure on both sides of the screen, alternating with the smaller 1:1.5 "V" figure in the shape-and-area-change stream and the smaller 1:3 "V" figure in the area-only-change stream. The other half of the infants saw the smaller 1:1.5 "V" figure on both sides of the screen, alternating with the larger 1:3 "V" figure in the shape-and-area-change stream and the larger 1:1.5 "V" figure in the area-only-change stream.

Results

Primary Preregistered Analysis

The reliability of the two looking-time coders was high (Pearson $r = 0.96$). Unlike in Dillon et al. (2020), infants did not look significantly longer to the relative length-and-area-change stream compared to the area-only-change stream [$t(47) = 1.29$, $p = 0.205$, $d = 0.19$; $BF = 0.339$; Figure 2B].

Secondary Preregistered Analyses

As in Experiment 1, we first identified influential participants by calculating Cook's distance in a linear regression on infants' raw looking times to each stream with Change Type (relative length-and-area change or area-only change) as a fixed effect. The analysis identified five influential participants. We reran the primary analysis on the data after removing these influential participants, and our results were consistent with the primary analysis [$t(42) = 0.71$, $p = 0.481$, $d = 0.11$].

We next ran a mixed-model linear regression on infants' raw looking times with Change Type (relative length-and-area change or area-only change), Size (bigger "V" or smaller "V"), Block (1, 2, 3, or 4), and Gender as fixed effects, and Participant as a random-effects intercept. Consistent with the primary analysis, we found no significant effect of Change Type ($\beta = 0.88$, $p = 0.377$), indicating that infants did not look longer to the relative length-and-area-change stream compared to the area-only-change stream. There was no significant effect of Size ($\beta = 0.69$, $p = 0.557$) or Gender ($\beta = 1.86$, $p = 0.117$), but there was a significant effect of Block ($\beta = -2.27$, $p < 0.001$), with looking time decreasing across blocks. As in Experiment 1, we also conducted this regression including partial datasets from infants in the planned age range (we received three such datasets), and since these results were consistent with the primary analysis, they are reported in the **Supplementary Material**. Finally, we ran a mixed-model linear regression with the same variables in the Bayesian framework. It revealed results consistent with the hypothesis-testing framework, with an estimate of 0.88 s (95% CI: $-1.07 - 2.83$) for the effect of Change Type, an estimate of 0.69 s (95% CI: $-1.56 - 2.94$) for the effect of Size, an estimate of 1.86 s (95% CI: $-0.39 - 4.11$) for the effect of Gender, and an estimate of -2.27 s (95% CI: $-3.15 - -1.40$) for the effect of Block on infants' looking times. As in Experiment 1, moreover, we repeated the primary analysis considering only the first two blocks. This analysis also showed that infants did not look significantly longer to the relative length-and-area-change stream compared to the area-only-change stream [$t(47) = 0.82$, $p = 0.415$, $d = 0.19$].

Finally, to evaluate the effects of distraction on infants' performance, we repeated Experiment 2's primary analysis but this time included the "distracted" infants, who would have met the inclusion criteria from Experiment 1. As outlined above, this sample included an additional 42 infants, and with this larger group of infants ($N=90$) we still did not find evidence that infants looked significantly longer to the relative length-and-area-change stream compared to the area-only-change stream [$t(89)=1.88$, $p=0.063$, $d=0.20$; $BF=0.630$]. To examine whether the findings of our exploratory analysis of distraction from Experiment 1 generalized to Experiment 2, we used a Spearman correlation predicting looking times by the number of distracting events, as in Experiment 1, with the expanded sample of 90 participants. There was no correlation between the number of distracting events and infants' preference for the relative length-and-area-change stream ($r_s=-0.09$, $p=0.402$; $BF: 0.300$).

Exploratory Analysis

To complement the exploratory analyses from Experiment 1, we first directly compared the results from this experiment to those of Dillon et al. (2020), Experiment 2B, using an independent samples *t*-test. While the difference between the two experiments was not significant [$t(21)=1.97$, $p=0.062$, $d=0.64$], the effect size was medium-to-large and similar (indeed slightly larger) than the effect size characterizing the difference between Experiment 1 to Experiment 1B of Dillon et al. (2020), which did show a significant difference. As in Experiment 1, infants looked longer at the stimuli online compared to the laboratory [$t(47.82)=-9.78$, $p<0.001$, $d=2.14$], suggesting that they saw the stimuli for a long enough time to show the expected effect.

We next evaluated whether infants looked longer to one side of the screen and whether the number of distracting events led to differences in overall looking, not just longer looking to the relative length-and-area-change stream. A little over half of parents (29/48) held their child over their left shoulder, and a mixed-model linear regression with Change Type (relative length-and-area change or area-only change) and Side (left or right) revealed a significant effect of Side ($\beta=3.76$, $p=0.009$), with infants looking more to the right versus left side of the screen. Nevertheless, there was no Change Type X Side interaction ($\beta=0.20$, $p=0.921$), suggesting that infants did not look significantly longer at the right side of the screen, for example, when that side presented relative length-and-area changes versus area-only changes, consistent with our primary finding. Finally, while the number of distracting events did not negatively correlate with a preference for the relative length-and-area-change stream, it did positively correlate with infants' total looking time ($r_s=-0.34$, $p=0.001$).

DISCUSSION

Unlike in Dillon et al. (2020), in Experiment 2, we found no evidence that 7-month-old infants looked significantly longer to shape changes instantiated as changes in the relative lengths of the arms forming simple 2D "V" figures. These results are

consistent with Experiment 1, which also failed to find that infants could detect subtle shape changes in 2D closed figures when tested online in their home environment. Experiment 2's null finding emerged regardless of its strict criteria excluding a large number of infants who experienced more than one distracting event during the testing session. Unlike Experiment 1, moreover, we found no relation between the number of times that infants were distracted and their ability to detect shape changes. This finding suggests that other factors, instead of or in addition to distraction, may affect infants' performance in home versus laboratory settings.

GENERAL DISCUSSION

Two experiments on young infants' shape discrimination adapted for an unmonitored online testing platform did not reveal infants' sensitivities to shape information as had been revealed robustly in laboratory experiments. In particular, unlike in Dillon et al. (2020), we found no evidence that 7-month-old infants looked significantly longer to the shape changes in triangles (Experiment 1) or the relative length changes in "V" figures (Experiment 2) over and above changes in figure position, orientation, and size. While exploratory analyses in Experiment 1 suggested that infants' failure to detect these shape changes might be due to their distraction, planned analyses in Experiment 2 found no relation between infant distraction and their change detection. Our findings suggest that other factors, instead of or in addition to distraction, may have instead affected infants' performance when tested online.

One possible factor that may have limited infants' success is the stimuli's presentation on small, personal computer screens. For example, while Smith-Flores et al. (2021) found looking-time results with toddlers tested online that were largely consistent with laboratory-based results, they speculated that their one null-finding – that infants failed to look longer at events in which an object appeared to move through another object after rolling down a ramp – may have been due to the events' being presented on a small screen, which minimized the visibility and salience of the violating object's trajectory. Similarly, the small screens used in the present study may have limited the visual saliency of the subtle shape changes. Indeed, the use of small screens in such cases may affect infants' performance whether or not they are tested online. Follow-up studies presenting different kinds and magnitudes of spatial information conducted in the laboratory using small screens may begin to address this possibility.

Among other developmental paradigms using looking time, moreover, change detection, in particular, relies on conveying that there are two discrete events being presented, with only one event displaying uniquely changing information that would draw infants' attention. Small screens may make this important aspect of the change-detection paradigm more difficult to convey, especially compared to contexts in which change-detection displays are presented on specialized equipment, like large projector screens or two separate monitors, as had been done in most laboratory studies. While the change-detection paradigm may have seemed ideal for adaptation to online

testing, in particular, because of its ability to yield reliable individual differences in longitudinal studies and its use of fixed duration trials, it may not be adaptable to online contexts or even other laboratory or field contexts if the testing in those contexts relies on small screens. Future laboratory studies presenting the same display used in Dillon et al. (2020) but on small screens may further clarify the role of screen size in eliciting infants' change detection of shape information.

Some developmental paradigms used with young infants, even those that seem well-suited to the constraints and goals of online data collection, may thus not yield results consistent with laboratory results that rely on highly controlled settings and specialized equipment, such as large screens. Testing those paradigms online is a necessary first step in creating robust tools and expanding the space of inquiry for developmental science conducted online. As the present study suggests, moreover, such investigations may also suggest limits to developmental paradigms that are not specific to online testing but have not yet been recognized in the laboratory. Such findings thus allow us to further refine both sets of tools and better understand the contexts in which infants' abilities can be reliably and robustly measured.

DATA AVAILABILITY STATEMENT

Example stimuli, preregistrations, data, and analysis code are publicly available at: <https://osf.io/vvaw7/>.

ETHICS STATEMENT

Our use of human participants was approved by the Institutional Review Boards at Massachusetts Institute of Technology (MIT); cede agreement for multi-site research at MIT and Harvard

University) and at New York University. Informed consent was obtained from parents or legal guardians for infants' participation in this study as well as for the publication of any potentially identifiable images.

AUTHOR CONTRIBUTIONS

Both authors designed the study, implemented the study, analyzed the data, and wrote the paper.

FUNDING

This work was supported by a National Science Foundation CAREER Award (DRL-1845924; to MRD) and a Jacobs Foundation Early Career Fellowship (to MRD).

ACKNOWLEDGMENTS

We thank Elizabeth Spelke and Kim Scott for their help with the design and implementation of the experiments, Gustaf Gredebäck and Elin Schröder for an informative discussion about their use of the change-detection paradigm, and Ofelia Garcia, Holly Huey, Nicole Loncar, Eli Mitnick, and Shannon Yasuda for their help with data collection.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.734592/full#supplementary-material>

REFERENCES

- Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980
- Cohen, L. B., and Younger, B. A. (1984). Infant perception of angular relations. *Infant Behav. Dev.* 7, 37–47. doi: 10.1016/S0163-6383(84)80021-1
- Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating Amazon's mechanical Turk as a tool for experimental Behavioral research. *PLoS One* 8:e57410. doi: 10.1371/journal.pone.0057410
- de Leeuw, J. R., Coenen, A., Markant, D., Martin, J. B., McDonnell, J. V., and Gureckis, T. M. (2014). "Online Experiments using jsPsych, psiTurk, and Amazon Mechanical Turk." in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, July 23–26. Available at: <http://psiturk.org/ee>
- Dillon, M. R., Izard, V., and Spelke, E. S. (2020). Infants' sensitivity to shape changes in 2D visual forms. *Infancy* 25, 618–639. doi: 10.1111/infa.12343
- Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., et al. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behav. Res. Methods* 48, 829–842. doi: 10.3758/s13428-015-0642-8
- Lauer, J. E., and Lourenco, S. F. (2016). Spatial processing in infancy predicts both spatial and mathematical aptitude in childhood. *Psychol. Sci.* 27, 1291–1298. doi: 10.1177/0956797616655977
- Lauer, J. E., Udelson, H. B., Jeon, S. O., and Lourenco, S. F. (2015). An early sex difference in the relation between mental rotation and object preference. *Front. Psychol.* 6:558. doi: 10.3389/fpsyg.2015.00558
- Leshin, R., Leslie, S.-J., and Rhodes, M. (2020). Does it matter how we speak About social kinds? A large, pre-registered, online experimental study of how language shapes the development of essentialist beliefs. *PsyArXiv* doi:10.31234/osf.io/nb6ys [Preprint].
- Libertus, M. E., and Brannon, E. M. (2010). Stable individual differences in number discrimination in infancy. *Dev. Sci.* 13, 900–906. doi: 10.1111/j.1467-7687.2009.00948.x
- Lo, C. H., Mani, N., Kartushina, N., Mayor, J., and Hermes, J. (2021). E-Babylab: an open-source browser-based tool for unmoderated online developmental studies. *PsyArXiv* doi:10.31234/OSF.IO/U73SY [Preprint].
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., and Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra: Psychology* 6. doi: 10.1525/collabra.17213
- Quinn, P. C., and Eimas, P. D. (1997). A reexamination of the perceptual-to-conceptual shift in mental representations. *Rev. Gen. Psychol.* 1, 271–287. doi: 10.1037/1089-2680.1.3.271
- Quinn, P. C., Slater, A. M., Brown, E., and Hayes, R. A. (2001). Developmental change in form categorization in early infancy. *Br. J. Dev. Psychol.* 19, 207–218. doi: 10.1348/026151001166038
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751

- Ross-Sheehy, S., Oakes, L. M., and Luck, S. J. (2003). The development of visual short-term memory capacity in infants. *Child Dev.* 74, 1807–1822. doi: 10.1046/j.1467-8624.2003.00639.x
- Sauter, M., Draschkow, D., and Mack, W. (2020). Building, hosting and recruiting: a brief introduction to running behavioral experiments online. *Brain Sci.* 10, 1–11. doi: 10.3390/BRAINSCI10040251
- Schröder, E., Gredebäck, G., Gunnarsson, J., and Lindskog, M. (2020). Play enhances visual form perception in infancy-an active training study. *Dev. Sci.* 23:e12923. doi: 10.1111/desc.12923
- Schwartz, M., and Day, R. H. (1979). Visual shape perception in early infancy. *Monogr. Soc. Res. Child Dev.* 44, 1–63. doi: 10.2307/1165963
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (part 2): assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/OPMI_a_00001
- Scott, K., and Schulz, L. (2017). Lookit (part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Sheskin, M., and Keil, F. (2018). TheChildLab.com a video chat platform for developmental research. *PsyArXiv* doi:10.31234/osf.io/rn7w5 [Preprint].
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to Foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Slater, A., Mattock, A., Brown, E., and Brenner, J. G. (1991). Form perception at birth: Cohen and Younger (1984) revisited. *J. Exp. Child Psychol.* 51, 395–406. doi: 10.1016/0022-0965(91)90084-6
- Smith, L. B. (2009). From fragments to geometric shape. *Curr. Dir. Psychol. Sci.* 18, 290–294. doi: 10.1111/j.1467-8721.2009.01654.x
- Smith-Flores, A. S., Perez, J., Zhang, M. H., and Feigenson, L. (2021). Online measures of looking and learning in infancy. *PsyArXiv* doi:10.31234/OSF.IO/TDBNH [Preprint].
- Starr, A., Libertus, M. E., and Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proc. Natl. Acad. Sci. U. S. A.* 110, 18116–18120. doi: 10.1073/pnas.1302751110
- Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., and Wagenmakers, E. (2019). A tutorial on Bayes factor design analysis with informed priors. *Behav. Res. Methods* 51, 1042–1058. doi: 10.3758/s13428-018-01189-8
- Verdine, B. N., Golinkoff, R. M., Hirsh-Pasek, K., and Newcombe, N. S. (2017). Spatial skills, their development, and their links to mathematics. *Monogr. Soc. Res. Child Dev.* 82, 7–30. doi: 10.1111/mono.12280
- Yu, C., and Smith, L. B. (2016). The social origins of sustained attention in one-year-old human infants. *Curr. Biol.* 26, 1235–1240. doi: 10.1016/j.cub.2016.03.026
- Zaadnoordijk, L., Buckler, H., Cusack, R., Tsuji, S., and Bergmann, C. (2021). A global perspective on testing infants online: introducing ManyBabies-AtHome. *PsyArXiv* doi:10.31234/osf.io/cnwh5 [Preprint].

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Bochynska and Dillon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Disruption Leads to Methodological and Analytic Innovation in Developmental Sciences: Recommendations for Remote Administration and Dealing With Messy Data

Sheila Krogh-Jespersen^{1,2*}, Leigha A. MacNeill^{1,2}, Erica L. Anderson², Hannah E. Stroup^{1,2}, Emily M. Harriott^{1,2}, Ewa Gut^{1,2}, Abigail Blum^{1,2}, Elveena Fareedi^{1,2}, Kaitlyn M. Fredian^{1,2}, Stephanie L. Wert^{1,2}, Lauren S. Wakschlag^{1,2} and Elizabeth S. Norton^{1,2,3}

OPEN ACCESS

Edited by:

Sho Tsuji,
The University of Tokyo, Japan

Reviewed by:

Natalie Ann Munro,
The University of Sydney, Australia
Przemysław Tomalski,
Polish Academy of Sciences, Poland

*Correspondence:

Sheila Krogh-Jespersen
sheilakj@northwestern.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 28 June 2021

Accepted: 22 November 2021

Published: 04 January 2022

Citation:

Krogh-Jespersen S, MacNeill LA, Anderson EL, Stroup HE, Harriott EM, Gut E, Blum A, Fareedi E, Fredian KM, Wert SL, Wakschlag LS and Norton ES (2022) Disruption Leads to Methodological and Analytic Innovation in Developmental Sciences: Recommendations for Remote Administration and Dealing With Messy Data. *Front. Psychol.* 12:732312. doi: 10.3389/fpsyg.2021.732312

¹Department of Medical Social Sciences, Feinberg School of Medicine, Northwestern University, Chicago, IL, United States, ²Institute for Innovations in Developmental Sciences, Northwestern University, Chicago, IL, United States, ³Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL, United States

The COVID-19 pandemic has impacted data collection for longitudinal studies in developmental sciences to an immeasurable extent. Restrictions on conducting in-person standardized assessments have led to disruptive innovation, in which novel methods are applied to increase participant engagement. Here, we focus on remote administration of behavioral assessment. We argue that these innovations in remote assessment should become part of the new standard protocol in developmental sciences to facilitate data collection in populations that may be hard to reach or engage due to burdensome requirements (e.g., multiple in-person assessments). We present a series of adaptations to developmental assessments (e.g., Mullen) and a detailed discussion of data analytic approaches to be applied in the less-than-ideal circumstances encountered during the pandemic-related shutdown (i.e., missing or messy data). Ultimately, these remote approaches actually strengthen the ability to gain insight into developmental populations and foster pragmatic innovation that should result in enduring change.

Keywords: developmental methods, remote adaptation, innovation, telepractice, analytic processes, COVID

INTRODUCTION

Child development is characterized by rapid transitions in social-emotional, cognitive, communicative, and motor abilities in the first years of life that are heavily influenced by the environment. Increasingly, the developmental sciences incorporate multi-level methods to most effectively capture intra- and inter-individual differences in developmental pathways. A typical research design utilized in developmental sciences is the longitudinal study in which participants are recruited at a young age, potentially even before birth, and followed across a pre-determined time series in order to gain a rich characterization of their development within a cohort.

Study visits may be multiple hours in length, involve intensive measurement (e.g., neuroimaging methods and behavioral observation) and are typically administered within controlled laboratory environments. Given the importance of comprehensive in-person assessment in the developmental sciences, the impact of the COVID-19 pandemic on research cannot be overstated. We argue, however, that this disruption to the status quo of developmental research has provided a unique opportunity for innovation, improved accessibility, and pragmatic application of many measures. The current paper discusses the adaptations to behavioral assessments that were adopted during the COVID-19 shutdown that allowed for continuous data collection in two longitudinal samples when in-person assessment was no longer possible.

Disruptive innovation is a concept commonly used in business and marketing to refer to the situation in which a novel technology, strategy, or model surpasses the current seemingly adequate version to attract a new audience or encourage the current audience to increase engagement (Christensen, 1997). Critical to this innovation is that it is not a novel discovery of a new product; for example, Amazon did not invent bookselling – they merely innovated the model for doing so online. Disruptions are not new, as Insel (2009) posited that mental health disorders are neurodevelopmentally unfolding syndromes, which was considered “disruptive insight” to the field of psychiatry. Critical to the theory of disruptive innovation is that established institutions have very little incentive to adopt the new model when the perception is that their current model (e.g., in-person assessment) is successful. Here, we argue that the COVID-19 pandemic has forced developmental science to engage in disruptive innovation in that the model for conducting assessments has changed and it should not return to the previous model once a healthy environment (or as close as possible) is restored.

The field of developmental psychology has been moving toward remote assessment *via* online methods in recent years. Lookit is an online platform in which caregivers can sign up to have their children engage in behavioral studies *via* webcam. Lookit studies typically assess visual attention *via* looking time and preferential looking (for a full review of Lookit, including strengths and limitations, please see Scott and Schulz, 2017a; Scott et al., 2017b). Caregivers can login at their convenience and their child can participate from home if a webcam is accessible to record their child's responses. Other platforms for unmoderated remote studies include Discoveries Online (Rhodes et al., 2020) and ChildrenHelpingScience.com. Some sites, including ChildrenHelpingScience.com, also provide the opportunity to sign up for appointments with study staff for an interactive experience remotely. Specific research labs have also set up sites, including TheChildLab.org and themusiclab.org, in which families can participate in remote assessments. A recent publication from Sheskin et al. (2020) takes the next step in proposing an online “superlab” to encourage reproducibility by sharing data across studies. Most, if not all, of these platforms support discovery-based experimental paradigms rather than adapt existing gold-standard measures. Here, we will take a focused look at the steps taken to adapt

standardized clinical assessments for remote administration *via* a moderated, interactive model.

An Illustrative Example

Illinois had a critical role in the early timeline of the pandemic spread given that the second identified case of COVID-19 in the United States was from a Chicago woman (Tribune Staff, 2020), and she was involved in the first person-to-person transmission in late January 2020. In mid-March of 2020, all non-essential in-person activities at Northwestern University were suspended, leading many research teams toward a remote work model. This was disruptive to the multitude of developmental science labs located at Northwestern. For the current paper, we will focus on two longitudinal studies being conducted within the Developmental Mechanisms Program at Feinberg School of Medicine and the Institute for Innovations in Developmental Sciences. The NIH-funded “When to Worry” (W2W) studies aim to jointly characterize markers of mental health and language disorder risk across the toddler period. Enrollment is ongoing and the majority of the sample was between 2–3 years of age at the time of the shutdown. Families could participate in the W2W studies if the participating child was between 12 and 18 months for the initial recruitment sample or 24 months of age with a language delay in the language delay sample. One biological parent completed surveys, therefore, eligibility included whether English was spoken at home at least 50% of the time (80% for language delay sample inclusion). The only exclusion criteria were a diagnosis of a developmental or congenital physical disability, or birth before 36 weeks gestation. The Promoting Healthy Brains Project (PHBP) is a randomized clinical trial aiming to use precision medicine approaches to tailor a maternal stress reduction intervention guided by biobehavioral real-time indicators. Women and their infants are followed until the child is 2 years of age. PHBP was enrolling pregnant women in March of 2020. Inclusion criteria include having a gestational age below 22 weeks, planned delivery at a Northwestern-affiliated hospital, ability to complete surveys, assessments, and intervention sessions in English, and specific technology criteria related to the delivery of a prenatal stress intervention (i.e., access to Wi-Fi and a smartphone). Exclusion criteria included pregnancy complications that place infants at risk of neurological disorders or a diagnosis of a severe chromosomal or congenital abnormality in the infant.

Both longitudinal studies included standardized behavioral assessments of the child, computer/tablet-based tasks, parent-child interactions, parent interviews, MRI, EEG, eye-tracking, and parent surveys at various timepoints. While some data collection, for example, survey administration *via* online tools such as REDCap, could continue unaffected, the move to remote working environments jeopardized the ability to collect in-person data, which was central to these studies. We rapidly adapted two behavioral assessments utilized to characterize children's development for remote administration. The Mullen Scales of Early Learning (Mullen, 1995) is a standardized behavioral assessment of children's development appropriate from birth to 68 months assessing the following domains: receptive language, expressive language, gross motor, fine motor, and visual reception.

The Mullen utilizes a series of prompts/activities often involving manipulatives to engage the child in the behavior of interest. This assessment is often administered in clinical research settings as it provides a *T* score centered on $M=50$, $SD=10$, which can be used to determine percentile rank and age equivalencies for each domain. The Mullen is designed to encourage the optimal performance of the child in a research/clinical setting. The Disruptive Behavior-Diagnostic Observation Schedule (DB-DOS; Wakschlag et al., 2008) is a research paradigm relying on behavioral observation that aims to “press” for irritable affect in infants and young children. The DB-DOS examines young children’s emotion regulation capacities and is a valid and reliable tool for distinguishing between normative behavior and clinically concerning disruptive behavior. This assessment utilizes a caregiver context in which the level of support provided by the caregiver on a given task is varied and an experimenter context in which the child interacts with a research assistant with little support from the caregiver. Tasks, often involving manipulatives, are designed to be developmentally appropriate to reflect everyday activities while still placing demands on young children’s regulatory capacities to elicit clinically salient behaviors (e.g., having to wait to play with attractive toys). The DB-DOS was initially designed for preschool-aged children, but a recently adapted version is available for toddlers ages 12-to 18-months. We will outline the steps our research team engaged in to adapt these measures to the stay-at-home orders related to COVID-19.

First, there was recognition that we were not the only ones scrambling. The significance of this pandemic and the related stay-at-home orders were affecting researchers at the global level. The Institute for Innovations in Developmental Sciences at Northwestern organized a sharegroup meeting that bridged developmental labs at both the biomedical and social sciences campuses. Developmental researchers and their teams from across disciplines were invited to present on their remote adaptations to research paradigms, as well as participate in discussions related to complicated issues associated with assumptions regarding participant population resources, including possession of computing equipment with cameras and audio, access to Wi-Fi connectivity, and whether caregivers would have the bandwidth to continue to engage in research at this stressful time. Outreach to Pearson (which publishes many standardized developmental assessments) led to qualified permission to adapt measures for remote administration, ensuring that researchers within this sharegroup were in compliance with legal contracts. This “stronger, together” mindset allowed the researchers to focus on adaptation while institutional resources could focus on regulatory and compliance issues and facilitate nimble transition to the remote environment.

We carefully adapted behavioral assessments to be pragmatic and engaging to meet the needs of our developmental populations while maintaining standardized practices, essentially converting a complex laboratory study into a field study (Glasgow and Riley, 2013; Morris, et al., 2020). Key to our use of the term pragmatic here is the reduction of a lengthy in-person visit to a more concise administration, and critical to success during the pandemic was that the administration could be conducted

remotely. The following sections include our recommendations for adaptation and implementation of remote behavioral assessments with infant and child participants (and their caregivers as test administrators/moderators). A central focus of these adaptations was careful consideration of issues related to scientific integrity, measurement validity, and construct continuity with in-person assessments before and after the pandemic.

The Move to Online: Technological Adaptation

Given that COVID-19 restrictions limited opportunities for in-person assessment in both the lab and home environment, the only alternative for data collection was to move online for remote protocol administration. Northwestern adopted Zoom software for online activities and our ongoing study activities were granted Institutional Review Board approval to collect data *via* this video conferencing platform. Zoom has numerous settings and our research team found the Zoom subreddit¹ to be instrumental to their awareness of updates and problem-solving issues. This section will discuss challenges and resolutions to issues related to the use of Zoom.

Early in remote protocol development, the research team developed PowerPoint presentations to present to participants on Zoom. The goal was to record children’s behavioral responses to prompts with visual stimuli (e.g., “Can you point to the ball?”). Immediately, the team encountered an obstacle related to screen recording as the Zoom default recording setting did not record the child’s face and instead recorded all audio and the screenshared PowerPoint presentation. To address this issue, the view in Zoom had to be set to “gallery” and the participant’s video set to “pinned” in order to record the child’s behavioral responses and not the screenshared PowerPoint presentation. The team also experienced issues screensharing with an app that was designed to assess executive function, as it was originally screenshared *via* Airplay, but that resulted in frequent audio and video lags. With assistance from the Northwestern’s IT department and Reddit, the research team amended protocols to include a third device (i.e., an iPad) from which to screenshare the app directly, no longer requiring Airplay. This troubleshooting was not limited to visual displays in Zoom.

Our team also had to adjust settings and use a third device to resolve audio issues within Zoom. The DB-DOS requires that an audio clip play during the “crying baby task.”² This audio clip of an infant crying is part of pretense that there is an infant off-screen who is in distress. The goal of this task is to measure how the child reacts to this stressor. The team utilized the “share computer audio” Zoom feature, which played the audio clip in a web browser open on the team’s computer. We found that the quality of the audio was not rendered perfectly *via* Zoom, although this may be due to differences in device speakers and families’ internet speeds. Zoom seemed to have particular trouble projecting the sound of a bell.

¹<http://reddit.com/r/Zoom>

²<https://www.youtube.com/watch?v=oL2B-AAAnsHo>

A ringing bell was used in several tasks as an indicator that time for a task had expired; however, Zoom often dampened this audio to filter out nonspeech sounds, ultimately preventing participants from hearing it. The team switched to the “cosmic” iPhone ringtone as it could be heard clearly *via* Zoom. Although this particular sound was effective for our purposes, we do not recommend the use of Zoom to convey audio information for tasks in which precise sound information is integral to the task’s purpose (e.g., sound discrimination or nonword repetition tasks).

In the beginning months of the pandemic shutdown, many caregivers struggled to use Zoom during remote visits. One critical role that research assistants took responsibility for was providing technological support during these visits, as the platform functioning was essential to ensuring fidelity of administration, accuracy of scoring, and child compliance with the tasks. One method for addressing this issue before a problem was evident was to discuss the caregivers’ comfort level with technology during the visit-scheduling phone calls (e.g., Manning et al., 2020). The research assistants provided support ahead of the visit regarding how to access the Zoom link and what kind of device/setup would be ideal to help alleviate caregiver anxiety and reduce troubleshooting in the moment of the visit. One issue we discovered was related to the participants’ view on their screen during the remote visit. Instead of the gallery view of speakers that is typically presented in Zoom, we wanted the participant to be able to see the PowerPoint slides that had been created to present stimuli (more detail provided below). As such, the research assistant must be cognizant of this issue and remind the caregiver to adjust the “speaker window” as needed to ensure the appropriate stimuli are in view. Young children also struggled when interacting *via* Zoom as they became confused or distracted by the research assistant or their own faces on the screen. These issues were minimized by turning off the participant view in the Zoom visit. We also found that this confusion decreased later in the pandemic, possibly because children and caregivers became more familiar with videoconferencing with family, school, work, and other contexts.

Preparation for Remote Administration

The first challenge to remote administration was ensuring that we could conduct standardized assessment within a setting in which we had less control (e.g., the participant’s home). First, we adapted our protocols and scheduling scripts to include information about the format of the remote visits. Some items or measures required very specific materials for the child to manipulate, such as a series of cups that nest inside each other, whereas other materials such as a spoon may be available in most homes (see **Supplementary Materials** for a list of generalized objects). Caregivers were contacted to ensure they were comfortable completing the visit at home and assisting in administration of some items. Many families in the study who had completed lab visits before the pandemic were accustomed to playing a role in measure administration, for example, the caregiver follows a series of written prompts in the caregiver context of the DB-DOS during typical

administration. Many caregivers expressed enjoyment in taking on the role of assisting with administering items and appreciated this as an alternative to visiting the lab for those activities.

An immediate challenge was ensuring the families had the necessary items to complete each task in the remote protocol. We created a visit box to send to each family that included the materials needed for the visit (see **Supplementary Materials**), varying slightly depending on the child’s age, preferences, and special considerations. Each package included a letter to the family explaining the materials in the box and provided more detail about the procedure for the visit, for example, instructions for the administration of the DB-DOS specifically outlining the caregiver’s involvement. In the box, materials for each assessment were placed in separate clear plastic zip-top bags and labeled with the test or activity, item number, and contents. The visit box was shipped to the participant’s home using 2-day shipping with tracking. When explaining the remote protocol, families were asked to avoid opening the visit box until the time of the visit to ensure that materials were not misplaced and to support children’s engagement with these novel items during the visit.

One challenge to the use of a visit box was the cost as it was not planned in our original study budgets. Materials accounted for a large cost as we determined that it would be best for families to keep the items in the visit boxes (e.g., crayons, bubbles, and small toys). Families expressed appreciation for this consideration of health and wellness, as well as the convenience of not needing to orchestrate return shipping. Our team also encountered issues with 2-day shipping, as the visit box was not always delivered within the specified timeframe. If families had not received their visit box, the remote assessment had to be rescheduled. We also had a number of visit boxes go missing during shipping, which meant delaying the study visit date even further to ship a new box. Overall, the visit boxes had some issues but allowed us to provide a standard set of materials to families in our studies.

Immediately prior to the remote study visit, we conducted a “home setup call” with each family. During this call, we asked whether caregivers had issues with internet access or Zoom that they would like to discuss. We asked the family to complete the visit in a small room with no toys present if possible, yet some home layouts did not allow for families to be in a separate room. We also asked that they set the visit for a time when other siblings or pets could be cared for, as study visits were often prolonged when there were multiple distractions present. We adjusted our protocols to build in breaks between assessments that required attention toward the computer screen (e.g., the Mullen) or that relied on caregiver-child engagement during frustrating tasks (i.e., the DB-DOS) so the participant had the opportunity to decompress. Although not every visit had the ideal setting, we were able to prepare the families for the structure of the protocol in advance and make changes to our structure to accommodate the real-world demands of the home study environment.

Although each remote adaptation was designed to be standardized across participants, there were obstacles that made this more challenging. First, the remote assessments were

designed to be administered on a computer with participants being recorded (for reliability scoring) *via* the computer's camera. The use of this technology assumes that the family has a computer in the home, which is not always the case. It was possible to administer the protocol on a smartphone, but this resulted in a decrease in the size of visual stimuli presented and greater difficulty for the research team to code the participant's responses (e.g., pointing) due to the smaller screen size. Notifications from the phone were sometimes a distraction during study visits. Additional technological limitations for some families included the cost of phone data or plans, varying internet speeds, technological expertise, and unreliable video quality. One possible future solution would be for the research team to loan a computer and/or a cellular hot-spot to a family for the purpose of the study visit. It is also important to consider the assumptions made when conducting remote study visits, including more generally whether the home is a safe place to conduct a visit and whether the research team is trained to respond appropriately if they identify reportable events when conducting a visit. A second assumption is that this protocol is easy for caregivers to administer: future research should examine the caregivers' perspective with regard to administering these measures with their child. Perhaps this first-person role in the evaluation of their child's abilities, including cognition, motor, and language skills, is not comfortable or "natural" for them. Understanding this perspective is essential as the field moves in this new direction. We will now discuss the adaptations we made to the Mullen and DB-DOS in more detail.

Adaptation of Assessment Protocols

All assessment adaptations were discussed in depth with the research team and piloted to determine whether the adaptation resulted in infant/child response data that could be coded prior to implementing the new protocols with research participants. The Mullen Scales of Early Learning is a standardized assessment of children's gross motor, visual reception, fine motor, expressive language, and receptive language abilities. To be clinically valid, it requires precise administration of a stimuli set with a trained administrator and the child in a controlled environment with very little distraction. As we have outlined above, this is not the ideal assessment for remote administration. However, this assessment was a primary outcome in the W2W study and therefore was critical to adapt for the home environment. The first step in adaptation was to examine the specific item administration for each domain of the Mullen to determine feasibility. We modified the in-person protocol to include which subtests and items were to be completed during the virtual visit as well as the administration order. Caregivers were presented with an introduction that discussed the domains of the Mullen and the expectations regarding their child's behavior (e.g., "This set of activities is designed to capture a wide range of skills that your child may or may not have just yet.") The introduction stressed that the assessment needed to be administered in a specific way and asked that the caregiver follow the instructions on the screen as closely as possible. Caregivers were also encouraged to praise their child regardless

of their child's response and to avoid using language like "correct" or "that's right."

We focused on 3 domains of the Mullen, listed in the planned order of administration: Receptive Language, Expressive Language, and Visual Reception. Two domains of the Mullen, Gross Motor and Fine Motor, were removed from the protocol due to time constraints and because we had the ability to gather information about motor development *via* online survey (Ages and Stages). Each domain was a separate PowerPoint to allow for flexibility in administration. Each Mullen item had their own slide(s) that could include an instruction prompt, a stimulus (e.g., an image from the Mullen Stimulus book), or a photo of Mullen materials (e.g., a ball, spoon, car, and chair). To reduce administration time and to optimize child compliance and attention, scores from the participant's Mullen that was conducted the previous year in-person were used to determine which item would start the domain; therefore, we assumed no regression in ability but ensured that children reached a basal. We focus on the Mullen for this section but note that our team made similar adaptations to the Bayley Scales of Infant and Toddler Development (4th edition) for the PHBP study. For this assessment, the Cognitive domain was excluded in its entirety from remote assessment due to complications with administration.

Some items could not be administered remotely and were removed from the assessment protocol. Decisions were made to exclude items when it was determined by a clinical assessment expert and our research team following evidence during piloting that the feasibility of instructing the caregiver to accurately administer the item was low (e.g., too many steps and complicated instructions). Additionally, many materials used in the Mullen are proprietary and we could not provide those to families. Although this was an obvious disadvantage as it relates to data collection, it was a necessity to ensure participant comfort and safety. To our knowledge, this is the first study reporting remote adaptations to standardized cognitive functioning assessments, such as the Mullen and the Bayley, resulting in little empirical guidance for how to produce standardized scores when items are missing. Therefore, raw scores will be used in most analyses. Non-standard assessments (i.e., when items are not administered) will be reviewed by a clinical assessment expert to determine their validity. Further, previous research has used clinically informed imputation methods for generating standardized scores when items are missing (McHenry et al., 2021). Using this approach, we will be able to generate standardized scores for research questions that warrant the standardization. Raw scores from the clinically informed imputed approach will be compared to the non-imputed scores before standardized scores are used in analyses.

Our remote administration protocols relied on screen shared PowerPoint slides that presented the assessment stimuli and prompts for the caregiver. These presentation slides were designed so that they are accurate, clear, consistent, and easy to read. One lesson from piloting was that confusion was reduced when one lengthy slide was divided into two shorter slides. For example, each Mullen item included a slide with instructions and a slide for the item administration that included any

necessary prompts and/or stimuli. When a Mullen item required the child to look at and/or point to a picture, the prompts for the caregiver were placed at the top of the PowerPoint slide just above the picture. The researchers had to take extra care to minimize the written instructions or cues for caregivers that offered additional hints to children. For example, one slide listed different colors that the caregiver should ask the child to identify (e.g., “point to the red crayon, point to the blue crayon...”) and the text of the colors matched the prompted color. The researchers realized that while the color coding may aid in clarity for the caregiver, it also provides a hint for the child. As such for this example, the text of the colors was changed to a uniform black. All instructions and prompts were displayed in a “user-friendly” manner, yet wording did not deviate from the Mullen manual. All prompts were typed in bolded font and presented within quotation marks. All actions (such as pointing) were typed in italics. Furthermore, text was consistently located in the same areas of each slide, so the parents were primed to read the instructions and prompts.

Whereas adaptations to the Mullen required that we adhere as strictly as possible to the standardized administration, we were able to adopt a more pragmatic approach when adapting the DB-DOS (Wakschlag et al., 2005, 2008). The DB-DOS is designed to elicit variability in behavioral and emotional (dys) regulation and to provide clinically informative ratings of irritability within the developmental context. Specifically, the DB-DOS uses “presses” to efficiently elicit typical/atypical distinctions in irritability in young children. Because of its objective to examine these patterns across interactional context, the DB-DOS includes presses that occur during interactions between the child with a caregiver and the child with an examiner. Naturalistic presses have ecological validity as they mimic those experienced in children’s daily lives (e.g., the child must wait while the caregiver is engaged in another task). We generated a broader, more flexible DB-DOS paradigm that had a number of pragmatic refinements that still retained essential features. We have termed this pragmatic adaptation of the DB-DOS, the Early Regulation in Context Assessment (ERICA). The ERICA has multiple modes of administration, can be employed beginning at birth, and may be coded *via* a single observation rather than through traditional multiple iterative video passes. Its core feature is the use of developmentally appropriate, ecologically valid presses retained from the DB-DOS, as these have been shown to elicit higher rates of variability than standard observations that do not include presses (Hampton et al., 2020).

To adapt the ERICA for remote administration, the paradigm was shortened from 45 min to 20 min. To achieve this, we prioritized tasks that included presses for multiple domains (e.g., frustration, irritability, and anger). As a meaningful interaction between a young child and examiner was difficult to construct remotely, only the caregiver context was included in the remote adaptation. Presses were adapted to include only tasks that required items feasible and not cost-prohibitive to send in the visit box (e.g., finger paints and bubbles). These pragmatic adaptations have resulted in an improved design for this established behavioral paradigm.

Finally, the research assistant and the caregiver had to work collaboratively over Zoom to administer assessments properly and to manage the child’s behavior. Caregivers were integral to the success of these remote visits, as they did the actual task administration with the child. Research assistants aimed to develop a strong rapport with the caregiver to ensure fidelity of task administration and standardization across families. Written and oral instructions were drafted and revised to ensure clarity while being mindful of maintaining a 6th-grade reading level. Research assistants engaged in partnership building strategies including acknowledging that the protocol could be difficult for the caregiver to administer, praising the caregiver’s effort in following instructed prompts, and emphasizing that the research assistant is available to help answer questions and to chat with the child if the caregiver needed a break. When caregivers showed hesitation or looked uncertain, pauses were enacted to ask if they had any questions regarding how to move forward. Research assistants reported that they felt it was important to meet the caregiver where they were most comfortable with regard to administration feedback. If a caregiver deviated significantly from the instructions (e.g., to the point that the task demand was now different), research assistants paused the task and provided gentle corrections and asked to have the item repeated, often with a slight delay. Deviations were noted in visit notes and flagged for review by a clinical assessment expert. While many caregivers welcomed corrections during administration, some become defensive or more nervous, which became an important area for feedback and growth during our training sessions. We also found that children often lost focus while waiting for the research assistant and caregiver to finish discussing instructions. In response, we implemented planned breaks for the child or we added small animations of animals to the PowerPoint slides during these transitions to keep them engaged. To ensure fidelity of administration and scoring, all assessments were recorded with caregiver permission and sessions were reviewed by a clinical assessment expert.

What We Lost and What We Gained: A Hybrid Approach

Unfortunately, some methods of data collection were not suitable for remote adaptation, specifically EEG, MRI, and eye-tracking. There are mobile versions of eye-tracking and EEG that were not feasible for our current studies given the shutdown restrictions. As restrictions regarding in-person activities lifted, the realization that we could return to the lab sparked a new focus: Can we optimize the protocol such that some of the study timepoints remained remote while additional new timepoints focused on these missed activities? Decisions had to be made about what was essential to addressing our programmatic research questions. Each study protocol was dissected to determine what assessments were not optimal for administration remotely. For the PHBP, two remote study visits were added: one when the participant was 7–9 months and a follow-up at 2 years. An original 12-month assessment timepoint was maintained with a new design: first, families complete a

remote study visit, followed by an in-person visit that includes MRI and EEG, as well as an abbreviated behavioral assessment that includes cognitive and executive function tasks that were difficult to administer online. The W2W study added a timepoint to measure parent-child interaction, parent stress levels, child language, COVID-19 illness, and the overall impact of the COVID-19 pandemic on families' everyday routines *via* videochat, and online surveys with support from a supplement from NIH. The inclusion of these additional timepoints was facilitated by supplemental grants that aimed to examine families' experiences during the pandemic. Here, the strength of the disruptive innovation is evident as the study design of incorporating both remote and in-person assessment facilitates rich characterization of families while reducing burden on them.

Our in-person protocols were also adapted to align with health recommendations from the CDC, capacity restrictions from Northwestern, and precautions necessary to keep our staff and participants safe. A pandemic research plan was drafted and submitted for approval by the Feinberg School of Medicine Office for Research. This included safety and health procedures, as well as occupancy limitations and scheduling accommodations to maintain the lowest level of health risk to our staff and participants. Some features of this plan include health screenings of the participants and the staff members conducting the visits 24 h prior to the in-person visit; temperature screenings upon arrival to the study visit; cleaning protocols and ventilation accommodations, including having a HEPA air filter in the study room; adult participants were required to mask and young children were encouraged to wear one throughout the study visit; and staff wore KN95 masks during all visits and were each provided with a face shield. With these safety procedures in place, we still faced hesitancy from participants to complete in-person visits. Some caregivers expressed reassurance in our safety procedures but did not feel comfortable having to take public transportation or ride-share. Additionally, many families faced issues with childcare due to other children being home. Pre-pandemic, we would provide families with childcare in the lab; however, this was eliminated due to capacity and staffing restrictions. For the families that did participate in-person, many caregivers expressed that this visit was one of the only excursions they had taken with their child since the pandemic began. As of June 11, 2021, Chicago has entered into Phase 5 opening (Illinois Department of Public Health, 2021), meaning that restrictions have been fully lifted in nearly all environments, including research settings outside of hospitals/clinics. As we move into this new level of comfort and an increase in in-person activity, it is important to reflect on how the ability to continue to collect data remotely was critical to characterizing the participants and their families during one of the most tumultuous times in recent history.

The move to remote study visits did allow for some opportunities and advantages. For example, *via* remote visits, we could continue to include families that moved out of state during 2019–2020. Previously, their participation would have been limited to survey and phone interviews because most would not be able to travel to the lab for in-person assessment

(our study did not budget for long-distance travel). Caregivers commented that it was easier for them to schedule the study visits because of the lack of commute time and the ability to conduct the assessment in their home. Also, providing this remote option helped us gain insight into the development of the child when caregivers were hesitant to come in-person. Children also appeared more comfortable during the remote assessment, possibly due to the familiar setting (e.g., their own snacks to eat and their own bathroom to take bathroom breaks). Whether this comfort then allowed the children to perform at a level that is a more accurate reflection of their skills and knowledge on standardized assessment is an open question for future research. Many of our in-person assessments relied on caregiver-child interaction (e.g., the DB-DOS); as such, the fidelity of those assessments was largely maintained. Standardized assessment, like the Mullen, presented unique challenges, as discussed. We highly recommend video recording of remote assessments, if possible, as this affords the opportunity to ensure fidelity of task administration and scoring *via* review.

Considerations and Strategies for Handling Missing and Messy Data

Methodological approaches to managing missing data are particularly critical for longitudinal research, as attrition is bound to occur. Although missing data are indeed commonplace in developmental studies, ignoring their presence and impact on study findings can lead to biased results and conclusions (Little and Rubin, 2002; Schafer and Graham, 2002; Jeličić et al., 2009). Arguably, the COVID-19 pandemic has fostered unavoidable and more extreme levels of missingness than what are typical (i.e., more than 50% missing; Enders, 2013), prompting creative problem-solving on the part of the researcher. Further, the pandemic may have introduced more measurement “messiness” or more measurement variability, including less standardization of assessments (e.g., distractions in the home) and collecting aspects of assessments in different ways (e.g., one Bayley scale was collected in person and another remote). In this section, we provide a brief conceptual overview on methods for handling missing and messy data, and practical steps we have taken in our own research for documenting and tracking missingness and changes in methods. We encourage researchers to seek out seminal papers on the topic for further information and guidance (Rubin, 1976; Little and Rubin, 2002; Schafer and Graham, 2002; Jeličić et al., 2009; Enders, 2013; Little et al., 2014).

Types of Missing Data

Although COVID-19 has exacerbated the issue of missing data in developmental research, these problems of missingness and messiness are not insurmountable. In fact, there are several robust methods for dealing with missing data that allow researchers to draw valid conclusions from the results. Before we determine the method for handling missing data, we must first identify the type of missing data with which we are working (i.e., why these data are missing). According to Rubin (1976), there are three common mechanisms for missing data.

Data are considered missing completely at random (MCAR) if the probability of missing data on a given variable is unrelated to the other measured variables. As an example, in order for our data on irritability to be considered MCAR, we would have to find that no measured variables in our study predicted whether an infant had missing irritability data. Data that are missing at random (MAR) are those that are related to variables other than the variable with missing data. In other words, data are MAR when the missingness is a result of other measured variables. Continuing with the same example, the data would be MNAR if missing rates for irritability were related to another variable in our study (e.g., harsh parenting), but not related to irritability. Finally, data are considered missing not at random (MNAR) when the probability of missing data on a variable potentially depends on the missing value itself. So using the current example, despite controlling for our other measured variables, infants high in irritability would be more likely to have missing values for irritability. Pauses in data collection due to the COVID-19 stay-at-home order may initially seem to be a source of MNAR, but participants with data gaps due to COVID-19 may not necessarily differ in a systematic way from those without these gaps (i.e., those individuals who visited the lab before the order was in place). Once the reason for missingness is assumed (given that some assumptions of MAR and MNAR are unable to be directly tested), researchers should report it in their manuscript, as well as the methodological rationale for handling the data (Enders, 2013).

Best Practices and Statistical Methods for Handling Missing and Messy Data

Documentation for Sensitivity Analyses

Documenting and tracking reasons for missingness in close proximity to the data collection process allows for more sophisticated missingness analyses once data collection is complete. Including questions about the status of data collection to track missingness and deviations from the original protocol can be used to derive variables for potential model parameters. Throughout the pandemic, we have recorded dates for suspension and resuming of in-person activity. From these data, we can construct a variable to differentiate participants who withdrew from the study from those who were physically prevented from providing data due to restrictions or government mandates on visits. Further, given our rapid response to changing method administration to continue collecting data, for any measures that vary in their mode of data collection (i.e., were administered remotely or in-person), we have created a field in our database to document which method applied to that individual visit. Sensitivity analyses can then be used to address these patterns of attrition and changing methods. As an example, we can examine whether scores on the Mullen vary by collection method (fully in-person vs. remote). First, we can create a variable for collection method by dummy coding the method of administration (e.g., 0=in-person; 1=remote). Then, we can use this variable to determine whether Mullen scores vary by collection method. If Mullen scores do not vary by collection method, then statistical analyses can proceed as planned. These

dummy-coded variables should also be considered for inclusion in the main study models as controls if there is theoretical justification (e.g., if the researcher would expect the outcome to change depending on method of collection). With respect to repeated measures data, we can take a missing modeling approach to test the most frequent occurrences of patterns of missingness. For example, we might find that missing the second measurement occasion is the most frequent type of pattern, or overall, we are finding five common patterns of missingness that apply to most of our sample. Again, we can dummy code these patterns and include them in a model. In a growth curve analysis, we can test whether missingness patterns affect the intercept or slope of our construct of interest over time. We may find that these patterns of missingness do not influence trajectories, and again, we can proceed as planned. Documenting dates during which measurement occasions occurred can also allow for a continuous time metric, for which we can model trajectories for the participants (D. Mroczek, E. Graham, & E. Beck, personal communication, December 09, 2020).

Statistical Methods

Multiple imputation and full information maximum likelihood (FIML) are two popular and robust methods for handling missing data that follow MCAR or MAR assumptions (Jeličić et al., 2009; Little et al., 2014), both of which we plan to leverage in our data analysis. Multiple imputation is the process of copying the original dataset to generate multiple datasets that fill in missing values with plausible estimates (Rubin, 1987). By using this method, the values are maintained in the datasets to prepare them for analysis. The analysis is then fitted on the imputed datasets and pooled estimates are derived. By creating multiple datasets, variability is increased and the findings are arguably more generalizable than if one were to rely on a single imputation (Jeličić et al., 2009). To produce this needed variation, 20 to 100 imputations are likely sufficient (Graham et al., 2007). Auxiliary variables, or those variables that are related to the variables with missing data, should be specified in the imputation to correct for some biases inherent to the nonresponse (Schafer, 1997). Multiple imputation methods are available in many statistical software programs.

FIML, by contrast, imputes missing data for deriving model estimates, but then deletes the imputed values after the analysis is complete. Thus, FIML will not produce a dataset with imputed values as multiple imputation does. FIML uses the data from partially completed variables to estimate parameters. In this way, linear relations between the missing data variable and the other variables in the model work to generate the estimates (Little and Rubin, 2002; Schafer and Graham, 2002). Many software packages are able to implement FIML, and for some modeling techniques, it is the default strategy (e.g., growth curve modeling; Enders, 2013). Both multiple imputation and FIML are widely used methods for managing missing data, but in some cases, one method may be preferred over another. For instance, FIML may be more appropriate when the dependent variable is incomplete, whereas multiple imputation does not distinguish between independent and dependent variables in the imputation process. FIML often requires that the distribution

of the variable with missing data be multivariate normal, whereas multiple imputation is less rigid (see Enders, 2013 for a review).

Although less frequently used, a Bayesian modeling approach can be applied for handling missing data. As mentioned, MAR and MNAR are assumptions and cannot be formally tested. Bayesian modeling can formalize these more subjective assumptions (Daniels and Hogan, 2008). With a Bayesian analysis, the imputation model and the analysis model are fitted at the same time, whereby estimates are acquired from posterior distributions of the parameters and missing variables (Ma and Chen, 2018). However, this approach is typically not recommended if one does not have prior experience with Bayesian modeling.

We have overviewed several potential methods for statistically handling missing data, but there are two notably flawed methods that should be avoided (Little and Rubin, 2002). Listwise deletion is the process of deleting cases that have missing values for all analyses, and pairwise deletion is the process of deleting cases depending on the analysis. Because both methods eliminate incomplete cases, the analysis has less power. Further, removal of cases because they are missing may introduce biases to the findings (Enders, 2013). Importantly, the appropriate method for handling missing data depends on the specific data and model in question. As mentioned, these methods require that the data meet several assumptions and depend on what percentage of the data are missing, causes of missingness, and patterns of missingness (Scheffer, 2002; Jeličić et al., 2009). The percentages of missing data for each study variable should be reported, regardless of which missing data method was used. Further, there may be added complexities with particular data types. For instance, researchers have debated how to handle missing neuroimaging data and whether and how these data should be imputed (Matta et al., 2018). However, when we properly track missing and messy data, we can learn to embrace the disruption that is so characteristic to our line of study. Rather than delete these cases, modern statistical approaches and thorough documentation can make up for lost ground and allow us to draw valid conclusions from our findings. We anticipate that we will be able to leverage multiple imputation and FIML techniques with the majority of our data.

Testing the Predictive Utility of Disruptive Innovative

An advantage of disruptive innovative is that we can test empirical questions about the predictive ability of our new methodology. A first question we can ask relates to the comparability of our methods, such as whether the remote version of our instrument measures the same underlying construct as the version performed in the lab. In the COVID-19 pandemic, it was not possible to collect both in-person and remote measures from each participant, hence the reason for the transition to remote assessment in the first place. However, given the innovation that has stemmed from these unprecedented circumstances, it would be valuable for future work to administer both versions of the measures to formally test their agreement.

Another question we would want to examine is whether our remote methods have predictive utility over more simplistic measures, such as surveys. For example, is it worth the burden to both the participant and the researcher to collect a remote measure of responsive parenting when a survey measure of responsive parenting might suffice? For parenting researchers, the resounding answer may be “yes,” but it is important to empirically test whether our remote measures hold predictive value for our outcomes of interest, particularly when measures may be more intensive. In a new study we have underway (Luby et al., 2019), we are seeking to answer this question by developing a risk calculator for generalizable risk prediction of preschool psychopathology. We argue that although multiple levels of analysis allow us to identify comprehensive risk for psychopathology, assessments at every level for every child may not be feasible and may be challenging to translate to real-world practice. Risk prediction algorithms, in particular, necessitate the inclusion of more intensive or burdensome measures when they add substantial value to the predictive model (Lloyd-Jones, 2010). The goal of our study is to test whether more cost- and resource-intensive measures (e.g., MRI, EEG, and behavior) have greater predictive utility of mental health prediction over less burdensome measures (e.g., survey). Further, the methods needed to predict mental health outcomes may depend on the level of risk for the individual child. For example, using the stoplight metaphor (Smith et al., 2018), children at high clinical risk (red) may receive immediate referral for treatment or prevention/intervention, children at low clinical risk (green) may only receive later testing at their regular well-child visit, and children with high clinical uncertainty (yellow) may require the more intensive measures to more accurately predict risk.

To empirically test the added value of these intensive measures, we can employ three key statistics: concordance (*c*) statistic, discrimination slope, and model calibration. The *c*-statistic is the most common statistic for discriminating risk calculator performance, representing the receiver operating characteristic curve (ROC) (AUC; D’Agostino et al., 1997). The AUC, ranging from 0 to 1, reflects the ability of the risk score to distinguish between having the disorder and not having the disorder. The discrimination slope indicates model improvement in sensitivity and specificity (Pencina et al., 2008). Lastly, calibration measures how closely the predicted probability aligns with real experience (D’Agostino et al., 1997). Using these statistics, we can determine whether a model including more intensive measures can better distinguish between disorder and no disorder than a survey-only model. In sum, by determining which indicators and methods are needed to best predict mental health, we can accelerate clinical translation to prevent disorder onset while limiting assessment burden for both the participant and the researcher.

DISCUSSION

Given this disruptive effect of the pandemic, what changes in developmental research are likely to endure in a “post-COVID” world? Here we argue, it should not be a return to “business

as usual.” While often through this adaptation process our research team felt as if there was no perfect solution, we did determine the optimal settings to conduct behavioral assessments with varying demands on the caregiver and child to support data collection during a global pandemic. Importantly, we plan to continue to use remote assessment protocols in future studies as we found this disruptive innovative to be critical to successful engagement with our research participants and see the potential for this to impact data collection more broadly in the field of developmental science.

Employing hybrid or fully remote research paradigms has great potential to improve representation in research. Typical, lab-based developmental science studies are more likely to engage Western, Educated, Industrialized, Rich, and Democratic (WEIRD) participants from a close geographic area, given that participation is often more accessible and convenient for such families (Sugden and Moulson, 2015; Nielsen et al., 2017). Because development is shaped by early experiences rooted in culture and other features of the environment (Greenough et al., 2002), it is unlikely that many developmental processes are truly invariant across sociodemographic and sociocultural groups, so engaging diverse participants is critical. Importantly, including diverse participants increases the generalizability of research findings (Hammer, 2011; Rowley and Camacho, 2015). In the current set of studies, English proficiency was an inclusion criterion for eligibility and all measures were designed for administration in English. One consideration is the requirement for caregivers to be literate in English given that instructions were provided in a letter included in the visit box and presented on the computer screen. When designing for remote administration, care should be taken to ensure that the demands of the tasks being administered comply with inclusion criteria and do not tax the participant excessively.

There are multiple reasons why diverse families may be less likely to engage and remain in traditional research studies, which new methods may address. Individuals from groups who have been disproportionately mistreated in research in the past, as is the case for Black Americans (Green et al., 1997), may have greater distrust of researchers and be less likely to engage in research, particularly in-person studies. For low-SES and urban caregivers, completion of study or intervention visits is hindered by availability of adequate transportation, child care, and timing of visits during working hours (Gross et al., 2001). Additionally, as we noted previously, we preferred presenting images on a computer screen compared to a smartphone screen for a number of reasons, including improving visibility. This is a limiting factor for participation, although we discuss the possibility of providing loaner computers, which should be considered when determining the feasibility of remote administration of measures. Beyond just recruitment and administration issues, retention for longitudinal studies with many visits over long periods of time can be more challenging for families facing economic hardship, due to frequent residential mobility and changes in contact information that may be more prevalent (Knight et al., 2009) and preclude study completion. Platforms such as Lookit (Scott and Schulz, 2017a) have transformed researchers’ ability to collect data that was previously

only possible in the lab. The benefits of offering remote studies that families can complete in their homes, at times convenient for them, may result in greater representation in research through increased opportunity for engagement for nearly all families.

Another important theme for developmental scientists to consider as we move past the pandemic is what research measures are “good enough” to answer the questions of interest (Blackwell et al., 2020; Morris et al., 2020). Whereas a study may have previously collected an in-depth lab-based assessment designed to measure a specific construct, the pandemic has forced researchers to reconsider whether a shortened, remote, or less burdensome method (e.g., a questionnaire) can fill that position (e.g., Manning et al., 2020). This will be an important theme moving forward, as what is most pragmatic or efficient has long been ignored in many developmental studies in favor of what is most in-depth. Pragmatic measures are certainly the future of developmental assessment, given the success of the National Institutes of Health (NIH) Toolbox with children and adults, and its upcoming extension version that covers infancy through early childhood. In its current version, assessment domains, including Language and Executive Function, require approximately 10 min each to administer, with scoring completed on an iPad.

Overall, the COVID-19 pandemic has led to disruptive innovation in methods for remote assessment that will transform research and practice for the better. Efforts range from reimaging and redeploying widely used measures, such as a “mobile” version of the NIH Toolbox (Weintraub et al., 2013), to researchers first considering designing studies to be remote assessment rather than defaulting to in-lab work. Remote data collection also allows unprecedented abilities to collaborate and collect data globally. It is our hope that the scientific and practical challenges that researchers faced during the pandemic will ultimately result in a field that is better equipped to address developmental science questions and provide innovative insights.

AUTHOR CONTRIBUTIONS

SK-J, LW, and EN were responsible for drafting and revising this manuscript and oversaw the research adaptations discussed, in collaboration with EA. LM contributed the analytic innovation section. HS, EH, EG, AB, EF, KF, and SW contributed to sections on the adaptations of the methods. All authors contributed to the article and approved the submitted version.

FUNDING

Promoting Healthy Brain Development *via* Prenatal Stress Reduction: An Innovative Precision Medicine RCT Approach (Lurie Children’s Hospital of Chicago; Wakschlag), R01MH107652 (Wakschlag), R01DC016273 and R01DC016273-A1S1 (Norton/Wakschlag), R34DA050266-S1

(Wakschlag). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ACKNOWLEDGMENTS

We acknowledge the assistance from Pearson in granting permission for our adaptations of their standardized assessments. We also thank Jessica Horowitz, Amy Biel, Alexandra Harpole, Aleksandra Wicko, and Emily Weinstein for their assistance

throughout the pandemic in helping to facilitate and enact these adaptations. Finally, we thank Erik Krogh-Jespersen who inspired the discussion of disruptive innovation and its application to remote assessment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.732312/full#supplementary-material>

REFERENCES

- Blackwell, C. K., Wakschlag, L., Krogh-Jespersen, S., Buss, K. A., Luby, J., Bevans, K., et al. (2020). Pragmatic health assessment in early childhood: The PROMIS® of developmentally based measurement for pediatric psychology. *J. Pediatr. Psychol.* 45, 311–318. doi: 10.1093/jpepsy/jsz094
- Christensen, C. M. (1997). *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Boston: Harvard Business School Press.
- D'Agostino, R. B., Griffith, J. L., Schmid, C. H., and Terrin, N. (1997). Measures for evaluating model performance. Paper presented at the Proceedings-American Statistical Association Biometrics Section.
- Daniels, M. J., and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton, FL: CRC Press.
- Enders, C. K. (2013). Dealing with missing data in developmental research. *Child Dev. Perspect.* 7, 27–31. doi: 10.1111/cdep.12008
- Glasgow, R. E., and Riley, W. T. (2013). Pragmatic measures: what they are and why we need them. *Am. J. Prev. Med.* 45, 237–243. doi: 10.1016/j.amepre.2013.03.010
- Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* 8, 206–213. doi: 10.1007/s11121-007-0070-9
- Green, B. L., Maisiak, R., Wang, M. Q., Britt, M. F., and Ebeling, N. (1997). Participation in health education, health promotion, and health research by African Americans: effects of the Tuskegee Syphilis Experiment. *J. Health Edu.* 28, 196–201.
- Greenough, W. T., Black, J. E., and Wallace, C. S. (2002). "Experience and brain development" in *Brain Development and Cognition: A Reader*. eds. M. H. Johnson, Y. Munakata and R. O. Gilmore (United States: Blackwell Publishing), 186–216.
- Gross, D., Julion, W., and Fogg, L. (2001). What motivates participation and dropout among low-income urban families of color in a prevention intervention? *Family Relations* 50, 246–254. doi: 10.1111/j.1741-3729.2001.00246.x
- Hammer, C. S. (2011). The importance of participant demographics. *American Journal of Speech-Language Pathology* 20, 261–261. doi: 10.1044/1058-0360(2011/ed-04)
- Hampton, L., Roberts, M., Anderson, E., Hobson, A., Kaat, A., Bishop, S., et al. (2020). What diagnostic observation can teach us about disruptive behavior in young children with autism. *J. Dev. Behav. Pediatr.* 42, 55–60. doi: 10.1097/DBP.0000000000000857
- Illinois Department of Public Health (2021). Phase 5: Illinois Restored. Available at: <https://coronavirus.illinois.gov/restore-illinois/phase-5.html> (Accessed June 18, 2021).
- Insel, T. R. (2009). Disruptive insights in psychiatry: transforming a clinical discipline. *J. Clin. Invest.* 119, 700–705. doi: 10.1172/JCI38832
- Jeličić, H., Phelps, E., and Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology* 45, 1195–1199. doi: 10.1037/a0015665
- Knight, G. P., Roosa, M. W., and Umaña-Taylor, A. J. (2009). *Studying Ethnic Minority and Economically Disadvantaged Populations: Methodological Challenges and Best Practices*. Washington, DC: American Psychological Association.
- Little, T. D., Jorgensen, T. D., Lang, K. M., and Moore, E. W. G. (2014). On the joys of missing data. *Journal of Pediatric Psychology* 39, 151–162. doi: 10.1093/jpepsy/jst048
- Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data. 2nd Edn*. Hoboken, NJ: Wiley-Interscience.
- Luby, J., Allen, N., Estabrook, R., Pine, D. S., Rogers, C., Krogh-Jespersen, S., et al. (2019). Mapping infant neurodevelopmental precursors of mental disorders: how synthetic cohorts & computational approaches can be used to enhance prediction of early childhood psychopathology. *Behav. Res. Ther.* 123:103484. doi: 10.1016/j.brat.2019.103484
- Lloyd-Jones, D. M. (2010). Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation* 121, 1768–1777. doi: 10.1161/CIRCULATIONAHA.109.849166
- Manning, B. L., Harpole, A., Harriott, E., Postolowicz, K., and Norton, E. S. (2020). Taking language samples home: feasibility, reliability and validity of child language samples conducted remotely with video chat versus in-person. *J. Speech Lang. Hear. Res.* 63, 3982–3990. doi: 10.1044/2020_JSLHR-20-00202
- Ma, Z., and Chen, G. (2018). Bayesian methods for dealing with missing data problems. *J. Kor. Stat. Soc.* 47, 297–313. doi: 10.1016/j.jkss.2018.03.002
- Matta, T. H., Flournoy, J. C., and Byrne, M. L. (2018). Making an unknown unknown a known unknown: Missing data in longitudinal neuroimaging studies. *Dev. Cognit. Neurosci.* 33, 83–98. doi: 10.1016/j.dcn.2017.10.001
- McHenry, M. S., Oyungu, E., Yang, Z., Hines, A. C., Ombitsa, A. R., Vreeman, R. C., et al. (2021). Cultural adaptation of the Bayley scales of infant and toddler development, for use in Kenyan children aged 18–36 months: a psychometric study. *Res. Develop. Dis.* 110:103837. doi: 10.1016/j.ridd.2020.103837
- Morris, A., Wakschlag, L., Krogh-Jespersen, S., Fox, N., Planalp, B., Perlman, S., et al. (2020). Principles for guiding the selection of early childhood neurodevelopmental risk and resilience measures: HEALthy brain and child development study as an exemplar. *Adv. Resilience Sci.* 1, 247–267. doi: 10.1007/s42844-020-00025-3
- Mullen, E. M. (1995). *Mullen Scales of Early Learning*. Circle Pines, MN: American Guidance Service.
- Nielsen, M., Haun, D., Kärtner, J., and Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *J. Exp. Child Psychol.* 162, 31–38. doi: 10.1016/j.jecp.2017.04.017
- Pencina, M. J., D'Agostino, R. B., Sr., D'Agostino, R. B., Jr., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat. Med.* 27, 157–172. doi: 10.1002/sim.2929
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63, 581–592. doi: 10.1093/biomet/63.3.581
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: Wiley.
- Rowley, S. J., and Camacho, T. C. (2015). Increasing diversity in cognitive developmental research: issues and solutions. *J. Cogn. Dev.* 16, 683–692. doi: 10.1080/15248372.2014.976224
- Scheffer, J. (2002). Dealing with missing data. *Res. Lett. Inf. Math. Sci.* 3, 153–160.
- Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman.
- Schafer, J. L. (2002). Missing data: our view of the state of the art. *Psychol. Methods* 7, 147–177. doi: 10.1037/1082-989X.7.2.147
- Scott, K., Chu, J., and Schulz, L. (2017b). Lookit (part 2): assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/OPMI_a_00001

- Scott, K. M., and Schulz, L. E. (2017a). Lookit (part 1): A new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/opmi_a_00002
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to Foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Smith, J. D., Berkel, C., Jordan, N., Atkins, D. C., Narayanan, S. S., Gallo, C., et al. (2018). An individually tailored family-centered intervention for pediatric obesity in primary care: Study protocol of a randomized type II hybrid effectiveness-implementation trial (Raising Healthy Children study). *Imp. Sci.* 13:11. doi: 10.1186/s13012-017-0697-2
- Sugden, N. A., and Moulson, M. C. (2015). Recruitment strategies should not be randomly selected: empirically improving recruitment success and diversity in developmental psychology research. *Front. Psychol.* 6:523. doi: 10.3389/fpsyg.2015.00523
- Tribune Staff. (2020). 6 months of COVID-19: Timeline of the outbreak and how politics, sports, entertainment and the economy changed. Available at: <http://www.chicagotribune.com> (Accessed September 15, 2020).
- Wakschlag, L. S., Briggs-Gowan, M., Hill, C., Danis, B., Leventhal, B., Keenan, K., Carter, A. et al. (2008). Observational assessment of preschool disruptive behavior, part II: validity of the disruptive behavior diagnostic observation schedule (DB-DOS). *J. Am. Acad. Child Adolesc. Psychiat.* 47, 632–641. doi:10.1097/CHI.0b013e31816c5c10
- Wakschlag, L., Leventhal, B., Briggs-Gowan, M., Danis, B., Keenan, K., Hill, C., et al. (2005). Defining the “disruptive” in preschool behavior: what diagnostic observation can teach us. *Clin. Child. Fam. Psychol. Rev.* 8, 183–201. doi: 10.1007/s10567-005-6664-5
- Weintraub, S., Bauer, P. J., Zelazo, P. D., Wallner-Allen, K., Dikmen, S. S., Heaton, R. K., et al. (2013). I. NIH toolbox cognition battery (CB): introduction and pediatric data. *Monogr. Soc. Res. Child Dev.* 78, 1–15. doi: 10.1111/mono.12031

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Krogh-Jespersen, MacNeill, Anderson, Stroup, Harriott, Gut, Blum, Fareedi, Fredian, Wert, Wakschlag and Norton. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Implementing Remote Developmental Research: A Case Study of a Randomized Controlled Trial Language Intervention During COVID-19

Ola Ozernov-Palchik^{1,2*†}, Halie A. Olson^{1*†}, Xochitl M. Arechiga¹, Hope Kentala¹, Jovita L. Solorio-Fieldier¹, Kimberly L. Wang¹, Yesi Camacho Torres¹, Natalie D. Gardino¹, Jeff R. Dieffenbach¹ and John D. E. Gabrieli^{1,2}

OPEN ACCESS

Edited by:

Dima Amso,
Brown University, United States

Reviewed by:

Lorijn Zaadnoordijk,
Trinity College Dublin, Ireland
Jonathan F. Kominsky,
Harvard Graduate School
of Education, United States

*Correspondence:

Ola Ozernov-Palchik
oozernov@mit.edu
Halie A. Olson
holson@mit.edu

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 01 July 2021

Accepted: 06 December 2021

Published: 07 January 2022

Citation:

Ozernov-Palchik O, Olson HA,
Arechiga XM, Kentala H,
Solorio-Fieldier JL, Wang KL,
Torres YC, Gardino ND,
Dieffenbach JR and Gabrieli JDE
(2022) Implementing Remote
Developmental Research: A Case
Study of a Randomized Controlled
Trial Language Intervention During
COVID-19.
Front. Psychol. 12:734375.
doi: 10.3389/fpsyg.2021.734375

¹ Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States, ² Harvard Graduate School of Education, Cambridge, MA, United States

Intervention studies with developmental samples are difficult to implement, in particular when targeting demographically diverse communities. Online studies have the potential to examine the efficacy of highly scalable interventions aimed at enhancing development, and to address some of the barriers faced by underrepresented communities for participating in developmental research. During the COVID-19 pandemic, we executed a fully remote randomized controlled trial (RCT) language intervention with third and fourth grade students ($N = 255$; age range 8.19–10.72 years, mean = 9.41, SD = 0.52) from diverse backgrounds across the United States. Using this as a case study, we discuss both challenges and solutions to conducting an intensive online intervention through the various phases of the study, including recruitment, data collection, and fidelity of intervention implementation. We provide comprehensive suggestions and takeaways, and conclude by summarizing some important tradeoffs for researchers interested in carrying out such studies.

Keywords: online studies, RCT, intervention research, developmental psychology, diversity

INTRODUCTION

Intervention Research in Developmental Science

One overarching goal of developmental research is to improve children's outcomes. The most direct way to achieve this goal is to implement an intervention – some manipulation of a child's experience or environment – and determine whether it leads to positive changes in outcomes. Not only do such studies allow researchers to test the efficacy of specific intervention programs, but they also play a crucial role in understanding developmental phenomena by elucidating causal mechanisms. A randomized controlled trial (RCT) design is a gold standard for establishing causality and efficacy in intervention research.

Despite the importance of intervention studies in developmental science, executing these studies is difficult. Because effect sizes tend to be small in developmental intervention studies, large samples are needed to detect significant effects (Lortie-Forgues and Inglis, 2019; Kraft, 2020). Interventions must be administered with high fidelity, which can be challenging at a large scale and when they require the involvement of

caregivers or educators (Fixen et al., 2005; O'Donnell, 2008; Barton and Fettig, 2013). While in-lab intervention studies allow for highly controlled testing environments, they run the risk of not generalizing to real-world settings (Lortie-Forgues and Inglis, 2019). Additionally, in order to substantially impact a child's experiences or environment, interventions typically have to be implemented over a long period of time (e.g., on the order of weeks to months). Both recruitment and retention of participants in developmental research intervention studies pose significant challenges.

Further, if interventions are to be translated into wide use, they have to be highly scalable to large numbers of children in diverse environments. In particular, the field of developmental research has recently come under scrutiny for predominantly studying WEIRD (western, educated, industrialized, rich, and democratic) populations (Nielsen et al., 2017). Even in the limited context of the United States, participants from lower socioeconomic status (SES) backgrounds are consistently underrepresented in research (Manz et al., 2010; Nicholson et al., 2011), and the majority of developmental science publications do not achieve a race/ethnicity distribution that matches that of the United States population (Bornstein et al., 2013). In addition to the profound issues related to equity (Lorenc et al., 2013; Veinot et al., 2018), lack of diversity and representativeness in developmental science threatens the generalizability of findings and fundamentally hinders our understanding of human development (Nielsen et al., 2017).

One major roadblock to the inclusion of more representative samples is the low participation rates of families from disadvantaged backgrounds in research (Heinrichs et al., 2005). There are multiple barriers to research participation that these families face, including informational barriers (not knowing about research opportunities), perceptual barriers (how families view the purpose and significance of research), and practical barriers such as lack of time and access to transportation (Heinrichs et al., 2005; Whittaker and Cowley, 2012). There are also many hard-to-reach communities in remote areas, far from universities and research centers. Practical barriers are most prohibitive for families from disadvantaged backgrounds (Lingwood et al., 2020).

Online Studies: New Opportunities for Developmental Intervention Research

Online developmental research studies are becoming increasingly popular and have advanced rapidly during the COVID-19 pandemic. The main benefit of online studies is that they allow families to participate in research from the convenience of their own homes. These studies can take multiple forms, including moderated/synchronous video-based studies (i.e., a live experimenter interacts with a child over a video conferencing platform, such as the Parent and Researcher Collaborative¹; see a review by Chuey et al., 2021), unmoderated/asynchronous video-based studies (i.e., through platforms that collect video without a live experimenter present, such as Lookit²; Scott and Schulz, 2017;

for review see Rhodes et al., 2020), and unmoderated app-based studies (Gillen et al., 2021). Despite the increasing popularity of online developmental research and the promise of these methods for increased diversity and scalability (Casler et al., 2013; Scott et al., 2017; Kizilcec et al., 2020; Rhodes et al., 2020), online intervention research is still very limited (but see Kizilcec et al., 2020 for an example).

There are multiple factors to weigh when deciding whether and how to implement an online intervention study. For example, moderated research studies – particularly ones that target underrepresented populations – require a large investment of resources and labor (Rhodes et al., 2020). Using an online platform may increase geographic and racial representation (Scott et al., 2017; Rhodes et al., 2020), but at a potential risk of excluding low-income participants due to a lack of reliable internet and technology (Lourenco and Tasimi, 2020; Van Dijk, 2020). Disparities in access to internet and devices – i.e., the “digital divide” (Van Dijk, 2020) – were particularly apparent early in the pandemic, and concerns were raised about whether online studies would inadvertently decrease diversity in developmental studies (Lourenco and Tasimi, 2020). Finally, implementing research studies in participants' homes, unlike in-lab studies, requires giving up some control over the study environment. In this paper, we describe some of the important factors to consider in the context of our experience implementing an intensive, fully remote RCT language intervention with third and fourth grade students (ages 8–10 years) from diverse backgrounds across the United States from summer 2020 – spring 2021. Notably, this study used a moderated online study design with extensive direct communication, and thus our suggestions are specific to this particular approach. We conclude by highlighting three main tradeoffs to think about when designing a remote intervention study with a developmental sample.

Case Study: A Remote Language Intervention Study During the COVID-19 Pandemic

During the COVID-19 pandemic, we implemented an RCT intervention to assess the impact of listening to audiobooks on reading and language skills. Third and fourth grade students were randomly assigned to the Scaffolding, Audiobooks-only, or Mindfulness (active control) group. Children in the Audiobooks-only condition received unlimited access to audiobooks *via* the Learning Ally platform³, curated based on their listening comprehension level. Children in the Scaffolding condition also received audiobooks and recommendations, as well as one-on-one online sessions with a learning facilitator twice per week, focused on improving their listening comprehension strategies and supporting their intervention adherence. The Mindfulness group completed a control intervention using a mindfulness app. The intervention period was 8 weeks for each group, with 2–3 h of pre-testing and 2–3 h of post-testing using a battery of measures administered *via* Zoom. We believe that this project will serve as an informative case study for other developmental researchers

¹<https://childrenhelpingscience.com>

²<https://lookit.mit.edu>

³<https://learningally.org/>

considering adapting intensive developmental interventions to an online format. Hypotheses, detailed methods, and results from the study will be presented in a separate manuscript (Olson et al., in preparation⁴).

RECRUITMENT

An important consideration for developmental researchers planning an online intervention study is whether they will be able to recruit a large enough sample size within a feasible time frame. Furthermore, researchers may be looking to recruit samples that are representative in terms of demographic variables like race/ethnicity and SES. As a case study, we will first describe our final sample characteristics, and then outline specific examples of recruitment efforts throughout the study period that led to this sample, including costs for various recruitment strategies.

Participants

Beginning in mid-summer 2020, we set out to recruit 240 third and fourth grade students (80 per group) with a broad range of demographic, geographic, reading level, and SES characteristics. To be eligible for the first pre-testing session, children had to be fluent in English, have a caregiver who spoke English or Spanish, and have no diagnosis of autism spectrum disorders or hearing impairments. Given that all sessions were held virtually, over Zoom, we unfortunately could not accommodate families who did not have internet or computer/tablet access ($N = 14$). However, because this study took place during the pandemic, many school systems provided children with access to these resources. We reached back out to families who expressed interest but initially lacked a computer and/or internet over the summer to see if they had been provided these resources by the school system during the school year. Since many families in poor and rural communities lack access to reliable internet (Lourenco and Tasimi, 2020; Van Dijk, 2020), our sample may not be representative of the most severely affected lower-income communities. Children were compensated \$20 per hour for all pre-testing and post-testing sessions (approximately 6 h total during the study). Caregivers were additionally compensated \$5 per survey for completing a total of ten surveys at the beginning and end of the study. Families also received lifetime access to the Learning Ally audiobook service after completion of the study, regardless of their group assignment.

Figure 1 shows demographic information for the 255 participants (age range 8.19–10.72 years, mean = 9.41, SD = 0.52) who were eligible for our study and were included in one of our three intervention groups, as well as how our sample compares to the United States Census data from 2020 (excludes participants who did not respond to these questions; NA = 24 for race/ethnicity, NA = 24 for maternal education, NA = 37 for paternal education). To demonstrate how the sample demographics in this study compare to similar in-lab and online

studies, we also show demographic distributions from three comparison studies (**Table 1** and **Figure 1**): a pre-pandemic longitudinal neuroimaging study conducted in our lab that relied on school partnerships and in-school testing for recruitment (Lab Study A, Ozernov-Palchik et al., 2017), a neuroimaging study conducted in our lab that used a combination of outreach events, advertisements, and social media to recruit participants (Lab Study B, Pollack et al., 2021), and an online intervention study conducted by another lab during the pandemic (Other Lab, Bambha and Casasola, 2021). We conducted a chi-square analysis to compare differences in the frequency of children with parental education of only high school between the current study and the four comparison samples (i.e., Lab Study A, Lab Study B, Other Lab, Census). The current study was not significantly different in the frequency of high school level education or below than the Lab Study A [$X^2(1) = 3.12$, $p = 0.078$] and Lab Study B [$X^2(1) = 0.3$, $p = 0.584$], but it had higher frequency of high school level education or below than the Other Lab study [$X^2(1) = 26.15$, $p < 0.001$] and lower frequency than the 2020 United States Census data [$X^2(1) = 76.6$, $p < 0.001$]. For a study conducted entirely online and during the pandemic, we successfully achieved a socioeconomically diverse sample comparable to pre-pandemic in-person studies that relied on in-school recruitment. Notably, the comparison online study – which did not specifically aim to recruit a diverse sample in terms of SES – included almost all mothers with at least a 4-year college degree. Thus, the transition to online studies does not automatically increase participant diversity in terms of SES.

We also evaluated differences in the frequency of white participants across the five samples. Our study had a lower frequency of white participants than Lab Study A [$X^2(1) = 13.58$, $p < 0.001$], Lab Study B [$X^2(1) = 35.19$, $p < 0.001$], the Other Lab study [$X^2(1) = 27.14$, $p < 0.001$], and the 2020 United States Census [$X^2(1) = 14.02$, $p < 0.001$]. The majority of developmental studies do not have representative samples in terms of racial diversity (Bornstein et al., 2013). There are important caveats to the comparison between the current study and the other lab studies, however. The in-lab studies were not conducted during a pandemic, and they involved neuroimaging. Despite their longitudinal nature, the in-lab studies did not include an intervention, which may have incentivized participation from some families. Nevertheless, although the comparison is not well-controlled, it suggests that we were successful in recruiting a diverse, representative sample of participants. Furthermore, we attained substantially more geographic diversity than is possible with in-lab studies. Our 255 participants came from a total of 26 states and 186 zip codes in the United States, plus Canada (**Figure 2**).

Overall Recruitment Strategies

To attain a diverse sample for our online intervention study, we tried several avenues for recruitment, including existing relationships with schools, new school partnerships, and online advertising. We received MIT Institutional Review Board (IRB) approval for all of our recruitment materials including flyers and social media ads in English and Spanish. These flyers and ads included a link directing caregivers to our participant screening

⁴Olson, H. A., Ozernov-Palchik, O., Arechiga, X. M., Wang, K. L., and Gabrieli, J. D. E. (in preparation). Effects of remote voluntary audiobook randomized controlled trial intervention on children's language skills. *Manuscript in Preparation*.

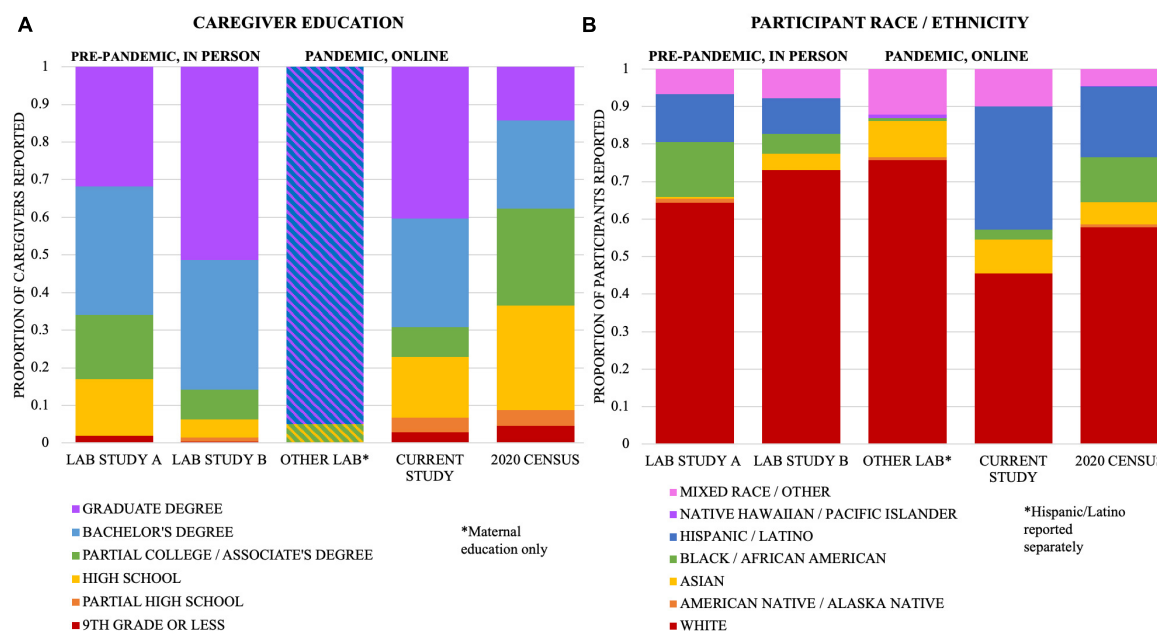


FIGURE 1 | Comparison to two studies from our lab conducted prior to the pandemic (Lab Study A, Ozernov-Palchik et al., 2017; Lab Study B, Pollack et al., 2021), one from another lab conducting a similar study during the pandemic (Other Lab; Bambha and Casasola, 2021), and the 2020 US Census. For Lab Study B, we included all participants who completed any portion of the study. **(A)** Highest level of parental education attainment, including both parents, for all who responded (Lab Study A, $N = 358$; Lab Study B, $N = 463$; Other Lab [maternal only], $N = 118$; Current Study, $N = 449$). 2020 Census includes all adults 25 years and older. **(B)** Parent-reported race/ethnicity of the child, for all who responded (Lab Study A, $N = 179$; Lab Study B, $N = 230$; Other Lab, $N = 115$; Current Study, $N = 231$). Participants who identify as Hispanic/Latino are counted in that category, regardless of race. Other categories reflect that race alone (not Hispanic/Latino). *Bambha and Casasola reported maternal education only: obtained high school degree (118/118), obtained 4-year college degree or above (112/118); and reported Hispanic/Latino separately from race (15/115 were Hispanic or Latino).

TABLE 1 | Comparison to three representative studies.

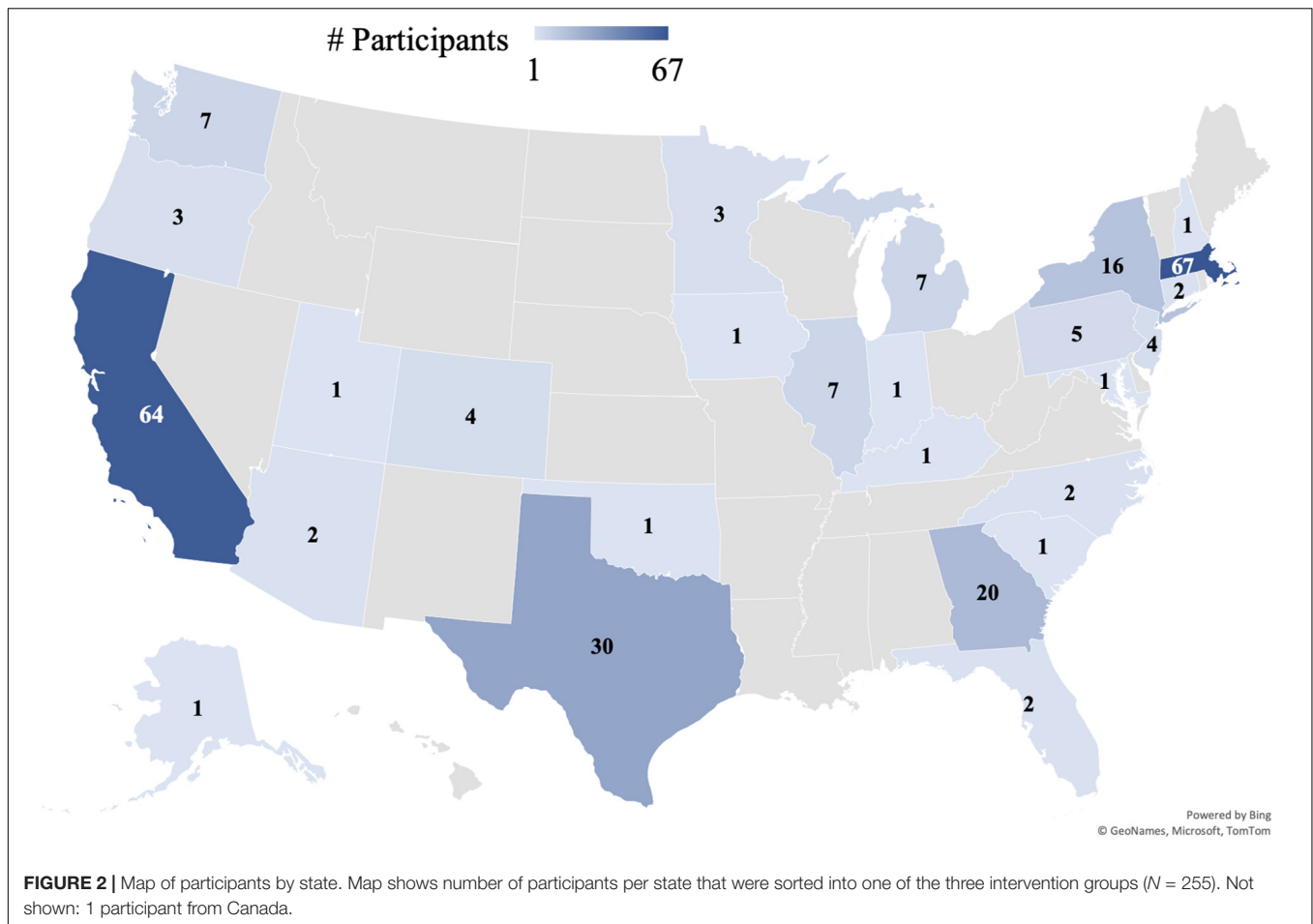
| Study | N | Age range | Setting | Recruitment | Time | Type |
|---------------|-----|------------|---------|--------------------------------|--------------|---------------------------|
| Lab Study A | 182 | 8–10 years | Lab | School partnership | Pre-pandemic | Neuroimaging/longitudinal |
| Lab Study B | 248 | 8–13 years | Lab | School outreach + social media | Pre-pandemic | Neuroimaging |
| Other Lab | 118 | 3–5 years | Online | Social media | Pandemic | Intervention |
| Current study | 255 | 8–10 years | Online | School outreach + social media | Pandemic | Intervention |

survey. All study data, including data from the screening survey, were managed using REDCap (Research Electronic Data Capture), a secure, web-based software platform designed to support data capture for research studies (Harris et al., 2009, 2019). The landing page, available in English and Spanish, briefly outlined the study and asked the parent or guardian to provide contact information, a simple demographic profile of their child, and other factors relevant to the study (e.g., access to technology). We included the question, “Does your child receive free or reduced lunch at school?” and prioritized contacting the families that responded ‘yes’ to this question. Below, we describe the efficacy of our different recruitment strategies, as well as our takeaways for other researchers considering these methods for an online intervention study.

School Partnerships

We began recruitment efforts in summer 2020 by reaching out to large and diverse school districts with whom we had

existing relationships. Our hope had been to disproportionately recruit lower SES students based on the profiles of the districts, such as public schools with high percentages of free/reduced lunch eligibility. We met with district leaders and principals, who expressed their enthusiasm and commitment to supporting our study. Fourteen schools, all with a large proportion of free/reduced lunch eligible families, officially partnered with our study. Outreach efforts by educators at our partner schools included pre-recorded phone calls to families, flyers, and text messages, with a range of 3–8 outreach attempts per school to their eligible students. This outreach yielded a relatively small fraction of the target number of students (Figure 3). It is important to note, however, that our school recruitment efforts took place during the early months of the pandemic when many educators were managing the logistics of school closures, and caregivers were getting accustomed to the new realities of remote learning. Additionally, our school partnership efforts were limited to schools with predominantly English



and Spanish speaking parents and caregivers, as we were not able to accommodate families in additional languages. Online intervention studies that choose to focus their recruitment on specific school districts should likewise consider the predominant language(s) spoken within the community, as we found that our study required substantial ongoing communication with families to provide appropriate support and ensure adherence (see “Family Communication and Retention” section, below).

Social Media

Our biggest recruitment success came from social media advertising through Facebook and Twitter. However, recruiting *via* these modalities introduced a unique set of challenges and considerations. One other online option we pursued was Craigslist targeted for specific zip codes, but this approach was ineffective due to Craigslist’s stringent policies regarding the categorization of ads.

Facebook

We first posted about our study on our lab’s Facebook page. Our lab had existing relationships with parent advocacy groups and other organizations that serve students with language-based learning disabilities. These organizations were more likely to include families from higher-SES backgrounds, so our initial

social media recruitment efforts were skewed toward this demographic. We then transitioned to paid Facebook ads. Our initial push was not as fruitful, primarily due to a low budget: we originally invested \$25 per posted ad, with each post spanning 3–5 consecutive days within a week. Each week, we launched a different ad until we exhausted our three differently themed ads (each available in English and Spanish), then started the sequence over again. After a month, we increased the budget to \$300 per posted ad for subsequent weeks. With this latter approach, we settled on three consecutive 24-h days, usually Friday–Monday. **Table 2** summarizes Facebook ad effectiveness for different representative configurations of ads.

Not surprisingly, it quickly became apparent that the amount of money invested resulted in increased study interest; the higher the investment, the more the ad is advertised across Facebook, Instagram, and Facebook messenger. The more the post is advertised, the greater the opportunity for engagement, and ultimately increased participation numbers. For future studies, if using Facebook, we recommend a generous social media budget to yield a large pool of participants. In total, we spent \$4,389 on Facebook advertisements over the course of the study, and a total of 131 of our 255 participants indicated that they found out about our study *via* Facebook (**Figure 3**), resulting in an average cost of approximately \$34 per participant

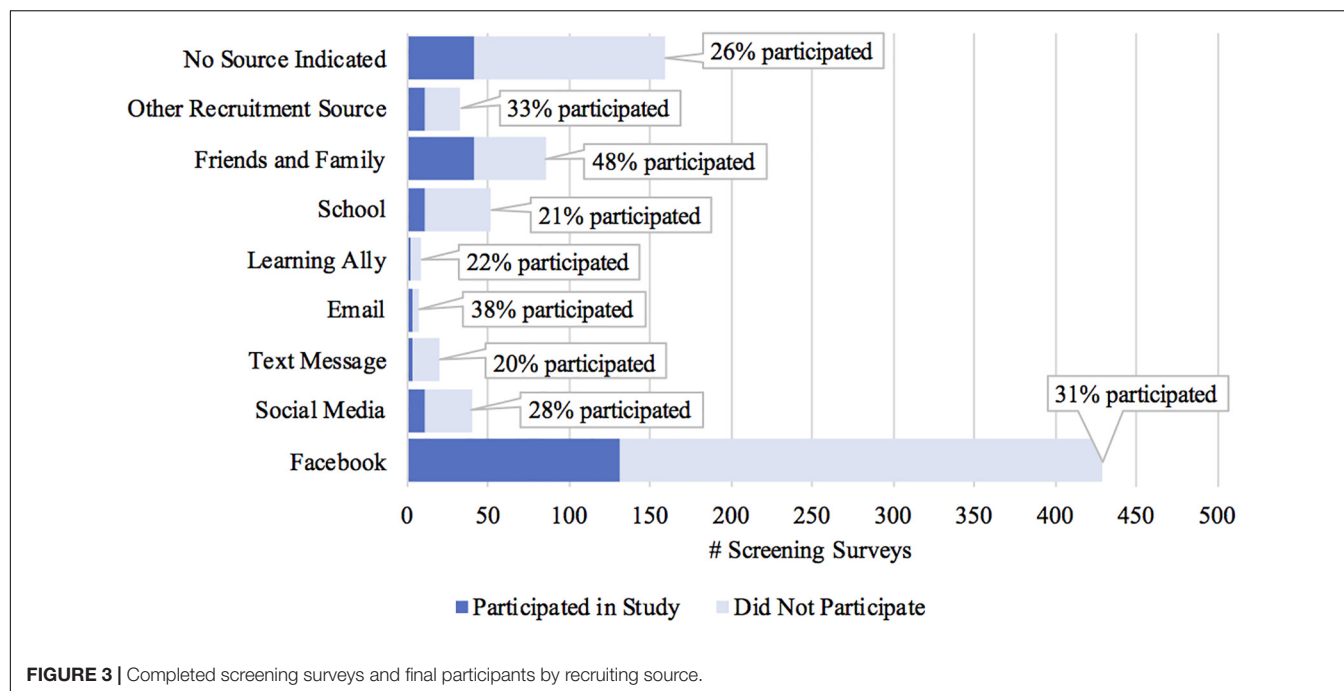


TABLE 2 | Effectiveness for three representative Facebook Ad configurations.

| Ad configuration | Total spend | Impressions | Clicks | Clicks per thousand impressions | Cost per click | Cost per participant |
|---|----------------|----------------|--------------|---------------------------------|----------------|----------------------|
| Set A: 25-mile radius around select cities | | | | | | |
| English Ads | \$1,714 | 273,448 | 3,030 | 11.1 | \$0.57 | n/a |
| Spanish Ads | \$363 | 78,593 | 709 | 9.0 | \$0.51 | n/a |
| English + Spanish Ads | \$2,077 | 352,041 | 3,739 | 10.6 | \$0.56 | \$17.02 |
| Set B: 10-mile radius around select cities | | | | | | |
| English Ads | \$1,089 | 160,038 | 1,823 | 11.4 | \$0.60 | n/a |
| Spanish Ads | \$373 | 61,952 | 524 | 8.5 | \$0.71 | n/a |
| English + Spanish Ads | \$1,463 | 221,990 | 2,347 | 10.6 | \$0.62 | \$86.05 |
| Set C: low SES zip codes | | | | | | |
| English Ads | \$579 | 99,180 | 579 | 5.8 | \$1.00 | n/a |
| Spanish Ads | \$271 | 37,984 | 212 | 5.6 | \$1.28 | n/a |
| English + Spanish Ads | \$849 | 137,164 | 791 | 5.8 | \$1.07 | \$283.06 |
| TOTAL | \$4,389 | 711,195 | 6,877 | 9.7 | \$0.64 | \$33.50 |

Total spend, number of advertisement impressions, number of clicks on our screening survey, number of clicks on our screening survey per thousand ad impressions, and the cost per click on our screening survey are shown for three of our Facebook advertisement campaigns. Estimated cost per participant was calculated based on participant report of how they found out about our study on the screening survey (N = 255 total participants began the intervention).

recruited *via* Facebook (Table 2). However, the actual cost per Facebook-recruited participant varied widely during different ad campaigns (Table 2).

To help us recruit participants from lower-SES backgrounds, we used targeted advertising. Facebook provides an option to target specific audiences by selecting cities, zip codes, educational level, age of child, individual interests, and more. While more individuals from targeted communities will see the post across their social media accounts, it does not necessarily mean that each individual who engages with the post will enroll in the study, so consistently posting is key to increasing enrollment rates. For instance, after boosting our recruitment success by targeting

ads at 25-mile radius circles around select cities (variously, Atlanta, Boston, Chicago, Dallas, Detroit, Houston, Los Angeles, Miami, New York, Philadelphia, Phoenix, and San Antonio), to target families closer to urban centers, we narrowed the radius to 10 miles in an attempt to recruit more lower-SES participants. Recruiting to this profile proved less successful than it was for the 25-mile radius group. We then used a “household income by zip code” list to try to further improve lower-SES recruitment, but as with the 10-mile radius effort, this approach was not successful. Table 2 shows estimated costs per participant (qualified and began the intervention) who learned about our study from one of the three ad campaigns. It should be noted that

these estimates rely on open-ended report of how participants learned about the study, and that these ad campaigns proceeded sequentially over different times during the year, with substantial variation in exactly which areas were targeted. Thus, while we think the estimates are informative for researchers considering these strategies, many factors likely influenced the number of participants we recruited.

Twitter

Learning Ally, the non-profit audiobook company that we partnered with for the study, advertised our study *via* Twitter (Table 3). We attribute their much higher ad engagement (about 10x what we saw with Facebook) to their large and strong following. This higher ad engagement did not translate to more sign-ups, however, as no participants explicitly identified Twitter as how they found out about our study.

Takeaways

School partnerships allow for greater control over participant demographics, as researchers can choose to partner with schools that have specific demographic profiles. However, establishing these partnerships takes time and effort, and may yield modest recruitment for an intensive, out-of-school intervention program. While it is certainly possible to establish school partnerships for an online intervention study, it does require substantial resources (both time and money) from the research team. Social media advertising brings the benefits of both large reach and precision targeting. Since online intervention studies do not have geographic constraints, this recruitment strategy may be beneficial for other developmental researchers considering implementing an online intervention.

Response rates per ad shown are quite small – close to 800,000 people viewing the ad yielded less than 150 actual participants. For the paid advertising, the cost ranged from \$0.25 to \$1.40 per ad click. This relatively wide difference reflects whether the audience knows the advertiser (in the case of Learning Ally's Twitter audience), how the ads were targeted by SES level (lower SES clicks had a higher cost), and what language the ads were in (English had a lower cost than Spanish). It is important to note that clicks do not remotely equate directly to study participants – the vast majority of people reaching the screener landing page (95%+) did not sign up for the study.

Overall, our recruitment efforts led to a representative sample of participants in terms of caregiver education and child's race/ethnicity (Figure 1). We also attained substantial geographic diversity, with participants from 186 different zip codes and 26 different states in the United States, plus Canada (Figure 2). Our sample was not substantially more diverse in terms of caregiver education compared to other studies run by our lab that aimed

to recruit diverse samples, but it was more diverse than another similar study run during the pandemic that did not explicitly aim to recruit a diverse sample based on caregiver education. Our sample was also more ethnically/racially diverse than similar in-lab studies and the general United States population. Thus, the transition to an online intervention format does not necessarily lead to more diverse samples on all dimensions without explicit efforts on those fronts, as well as a considerable recruitment budget.

FAMILY COMMUNICATION AND RETENTION

Another factor developmental researchers will need to consider when adapting to an online protocol for intervention studies is how to ensure continued engagement and adherence to the program. During our study, not only were we collecting data and administering an intervention online, but we were also doing so during a global pandemic. Families dealt with illness, death, financial stress, technological challenges, and other difficulties over the course of the study. We adapted our communication protocols to be as supportive to families as possible. We believe that these lessons are also worth sharing, as even in non-pandemic times, families encounter these and other challenges.

Personalized Communication Methods

Having robust procedures for family scheduling and communication was vital to our study. We had a dedicated research team whose primary role was to contact families and answer any questions that came up. This team included two full-time research staff, as well as 2–3 undergraduate research assistants available to troubleshoot specific questions regarding the use of the audiobook app. We received 15–35 emails per day regarding scheduling, rescheduling, payment requests, score report updates, app issues, etc.

Before the study began, we drafted email and text templates for key communication points at various stages before, during, and after the intervention. For example, we had templates for program orientation and onboarding procedures, appointment confirmation and session reminders, as well as periodic check-ins. In our screening form, we asked for each family's preferred method of communication, and we used this method throughout the study. To ensure consistency in communication, one researcher was assigned to each family and handled all communication for that family. While communicating with families using various methods (i.e., emails, phone calls, text messages) was more time and labor intensive, we found that it boosted participation throughout the duration of our study. We observed high retention rates overall, but there was still attrition (Figure 4). Text and email reminders helped minimize missed appointments. If participants missed a session or were generally more challenging to communicate with, we noted this for their next session and asked the tester to send an additional reminder the day of the testing session to ensure attendance.

For the Scaffolding Group in our study (i.e., the intervention group that met biweekly with 'learning facilitators' in addition

TABLE 3 | Effectiveness for Twitter Ads.

| Ad configuration | Total spend | Impressions | Clicks | Clicks per thousand impressions | Cost per click |
|------------------|-------------|-------------|--------|---------------------------------|----------------|
| Total campaign | ~\$450 | ~20,000 | 1,793 | 91.0 | \$0.25 |



FIGURE 4 | Participant pipeline and attrition.

to listening to audiobooks), the average family required approximately 37 points of contact throughout the study. This included appointment confirmations, reminders about reading

books, payment details, and parent surveys. Similar levels of communication were required for the other groups (i.e., Audiobook-only and Mindfulness), with around 24 points of

contact per family. Importantly, however, the number of contact points per family within each group varied based on families' circumstances. Families with limited access to and knowledge of technology at home required additional support throughout the study from our research team. Families with more variable work schedules were more likely to miss sessions or need to reschedule. Thus, we strongly advocate for clear, consistent, and individualized communication with all families, which may especially affect the enrollment and retention of the participants from disadvantaged backgrounds.

Importance of Bilingual Research Personnel

There was a large proportion of Spanish-speaking families in our partner schools. Our final sample included 12% Spanish-speaking participants (30/255), and we had two bilingual Spanish-speaking full-time researchers to support these families. At the beginning of the study, there was a large effort to translate all study materials, surveys, and additional resources into Spanish. Although most of the translation effort was front-loaded, there was still a need for Spanish-speaking researchers throughout the study for family communication.

Scheduling

Since we had families from across the United States participate in our study, we had to account for multiple time zones when scheduling sessions with testers and learning facilitators. We were able to schedule sessions around each family's schedule, including weekend and evening sessions. Each tester had a personal, secure Zoom link that was sent to the family before their scheduled session. Unlike in-person data collection, there was no limit to how many sessions we could book at one time, since physical space was not an issue. Testers called and attempted to troubleshoot with the family if the participant had difficulty getting onto Zoom. The child could complete sessions on a computer or tablet; we also allowed children to log onto the Zoom session *via* a cell phone in circumstances where no other option was available (only for tests without visual stimuli, as image size would be significantly reduced on a phone screen).

Retention

Most families who expressed initial interest by filling out our screening survey did not end up participating in our study. We experienced high attrition between screening, pre-testing, and group assignment. However, once participants completed onboarding procedures and began the 8-week intervention, attrition was quite low (**Figure 4**).

Flexible accommodations to families' individual needs minimized mid-study attrition, but it could not be entirely prevented. In some cases, children were very resistant to participating in the intervention. For example, given the new distance-learning protocols implemented during the COVID-19 pandemic, some children reported not wishing to have more screen-time. This may be relevant for future studies if educators continue to rely on screen-based technology for learning in and out of school. In instances where children were especially resistant to participating, we did not pressure them to continue. In other instances, however, families simply stopped responding

to emails and texts. We observed the greatest non-responsiveness at the end of the program, when attempting to schedule post-test sessions. When faced with non-responsive participants, we first followed-up with multiple (~5) reminder emails, phone calls, and/or text messages, then we issued one final check-in email before suspending any further attempts to reach out.

Takeaways

Overall, we credit the efficiency of our communication pipeline to the use of pre-drafted email/text templates and maintaining an active log of all communications. We recommend frequent and consistent communication with participants to minimize attrition when conducting large-scale online intervention studies. Being timely with responses encourages participants to continue with the study and increases their participation at the post-testing portion of the study. To manage a large number of participant questions, it is important to have a main contact person for each family. We found that regular interaction with our families, *via* their preferred mode of communication, was effective in establishing rapport and maximizing engagement. Moreover, it is critically important to have bilingual staff who can closely support families who may speak another language. Finally, detailed study orientation materials and clear, step-by-step onboarding procedures are useful to ensure that participants understand all study requirements and to preemptively troubleshoot potential barriers to participation.

DATA COLLECTION

For developmental researchers that typically utilize in-lab assessments, a major adjustment when transitioning to online intervention studies is adapting measures for online administration. Here, we describe the measures we used, how assessment scores compared to in-lab administration of the same assessments, how we dealt with variable testing environments, and how we trained our team to administer assessments online.

Behavioral Battery Adaptation

Adapting assessments for online administration required careful consideration to ensure feasibility for both testers and participants. We decided to administer all assessments over Zoom, which allowed testers to directly interact with participants in real-time. For scoring purposes, we audio- and video-recorded each session and stored these recordings securely. The Zoom platform enabled testers to share their screens, allowing us to display scans of stimulus items and online assessment platforms.

Online administration of the assessments in our battery required various considerations and adaptations (**Table 4**). Some tests had already been adapted for online administration, and we used the publisher's online administration and scoring platform. Other tests required tracking the child's responses and simultaneous scoring that was not viable *via* the computer. We mailed packets to each tester containing printouts of these assessment score sheets along with dry-erase markers and plastic protector sheets. This packet also included a copy of the testing manual containing the required materials and procedures for

all tests. These materials allowed testers to have fewer files open on their computer at once during the session. We used DropBox (MIT provides large storage space to its affiliates) to upload all materials for tester access (e.g., stimulus item scans, administration guidelines, etc.), and we used team Slack as a way to troubleshoot or to ask questions before, during, or after test administration.

Tester Training

The testing team consisted of graduate students from speech and language pathology or early education programs with experience administering psychoeducational evaluations to school-aged children. All testers were native English speakers, and some were also fluent in Spanish. All testers had prior knowledge of the Zoom platform and different file storing/sharing programs (e.g., DropBox, Google Drive). Testers were trained remotely on administering and scoring our assessment battery. Before starting their first session, testers scored a video-recorded session and were deemed ready if they achieved 95% reliability with the first scorer (an experienced tester). A team member reviewed and scored the video recording of each tester's first session with a child and gave them feedback as necessary. Training continued until the testers were able to administer and score all assessments with high accuracy. Testers were blind to participants' group assignments. One benefit of online testing is the ability to easily video record testing sessions. Doing so helped facilitate a more thorough reliability assurance than for in-lab studies that tend to only audio-record sessions.

Remote Administration

We also needed to adapt our general assessment administration procedures. Each session began with the tester confirming the child was in an optimal testing environment, and adjustments were made if necessary (i.e., moving to a quieter space in the home). Caregivers were asked for their permission to have the Zoom session recorded. The tester then reviewed the consent form with the caregiver and the assent form with the child, which had been emailed to the family before the session, and obtained verbal consent from both the caregiver and child. If the family's primary language was Spanish, the initial session was scheduled with a bilingual tester, or another bilingual member of the team joined the session to obtain consent in Spanish. Testers then administered the assessments. These were split across 2–3 sessions, as the battery of tests was extensive, and children generally fatigued after about 90 min. Immediately following the session, testers uploaded the recordings of both the verbal consent/assent and the testing session to a secure server, and submitted records of participants' responses.

Finally, we needed to establish data management and scoring procedures that ensured accuracy in the online setting. Since paper record forms could not be centrally stored with all of our testers working remotely, we created Google Forms to record participants' responses for most assessments. Having digital copies of item-level responses helped with easily calculating reliability for each assessment (Table 4). The Google Forms were used to generate spreadsheets of participant data for each assessment. All records only used participant

IDs. Other assessments required the use of the developer's platform for scoring.

Testing Environment

The testing team encountered a variety of challenges unique to the virtual testing environment. In-person assessment allows for more knowledge of and control over what participants are doing during the session. With online administration, we relied more on children and caregivers to achieve consistency in the testing environment. For instance, during the online sessions, we needed to make sure that participants could see and hear what we expected them to, despite not having direct control over the visual display and audio output of their devices. Thus, testers regularly asked participants to confirm that they could see the screen-shared materials and hear their voice clearly, adjusting the size of materials on display and asking children to adjust their speaker/headphone volume as necessary.

The most common issues were loud background noise in the home and poor internet connection, which often affected audio quality for the participant, tester, or both. It was sometimes difficult to judge the quality of what the child was hearing, especially when caregivers were not present to provide feedback. For assessments that involved timed performance or stimulus items that could not be repeated, testers made adjustments to reduce validity concerns. If there was background noise and the child did not have a quieter space, testers asked the child to put on headphones or saved listening tasks for the following session when the child might be in a quieter environment. Child responses were often difficult to discern when answer choices involved rhyming letters (e.g., A, B, C, D), even after asking the child to repeat the response. In these instances, testers requested that the child type their answers into the chat on Zoom.

Internet connectivity and other technical factors (e.g., the ability to download and play audio files provided by the team) varied widely across participants and between sessions. Sometimes testers turned off the video portion of the Zoom call in an attempt to improve the audio connection. The team also encountered minor technical issues with specific aspects of the online administration process, such as problems with using the "Remote Control" function on Zoom on certain types of computers.

At times, the participant's home environment was distracting for other reasons, such as family members or pets entering the room. Many children completed the testing from a desk, but many others completed it while sitting on a couch or in their bed, and some children needed reminders to sit up or change position to better focus on assessment tasks. Because some caregivers chose to remain in the room during testing, testers occasionally encountered caregivers who continued to help their child despite the tester's requests not to. In particular, because caregivers were often off-camera, it was sometimes difficult to gauge the extent of the support given by the caregiver. The presence of caregivers in the room may have made some children more self-conscious about their performance, whereas other children appeared comforted by their presence. Also, because the tester could not see the child's screen, some children may have attempted to look up answers to certain testing questions, though

TABLE 4 | Assessments and adaptations for remote administration.

| Assessment | Description | Adaptations | Sample reliability coefficients | Publisher reliability coefficients |
|---|---|---|---|---|
| Kaufman Brief Intelligence Test, 2nd Edition (KBIT-2) – Matrices ¹ | Standardized non-verbal IQ assessment | Scan of stimulus items screen-shared <i>via</i> Zoom | α : 0.83 Split-half: 0.81 | Split-half: 0.81–0.88 |
| Clinical Evaluation of Language Fundamentals, 5th Edition (CELF-5) – Understanding Spoken Paragraphs ² | Standardized test of listening comprehension | Administered <i>via</i> Zoom | α : 0.74 Split-half: 0.79 | α : 0.75–0.85 |
| Dynamic Indicators of Basic Early Literacy Skills (DIBELS) ³ <ul style="list-style-type: none"> • Word Reading Fluency (WRF) • Passage Reading Fluency (PRF) • Multiple Choice Reading Comprehension (MCRC) | Standardized measures to assess reading skills; MCRC is a computer-administered standardized test | WRF and PRF: Digital forms screen-shared <i>via</i> Zoom. Tester recorded errors on online progress monitoring site from publisher. MCRC: Tester screen-shared and child was given control of tester's screen to select multiple choice answers. Alternative was to have child orally tell tester which answer to select (when child was unable to utilize "Remote Control"). | Item level data was not available | |
| Peabody Picture Vocabulary Test, 5th Edition (PPVT-5) ⁴ | Standardized receptive vocabulary assessment | Images screen-shared <i>via</i> Zoom using publisher materials adapted for digital use (Q Global). | α : 0.96 Split-half: 0.96 | α : 0.97 |
| Wechsler Abbreviated Scale of Intelligence, 2nd Edition (WASI-II) – Vocabulary ⁵ | Standardized vocabulary assessment | Scan of stimulus items screen-shared <i>via</i> Zoom. | α : 0.8 Split-half: 0.82 | Split-half: 0.88–0.93 |
| Comprehensive Test of Phonological Processing, 2nd Edition (CTOPP-2) – Non-word Repetition, Memory for Digits, Blending Words ⁶ | Standardized measures to assess baseline working memory skills | Audio files sent to families to download ahead of time; child/caregiver asked to play each file from their computer during assessment. | NWR α : 0.73 Split-half: 0.76 MD α : 0.8 Split-half: 0.84 BW α : 0.84 Split-half: 0.86 | α : 0.77 α : 0.8 α : 0.8 |

¹Kaufman, 2004; ²Wiig et al., 2013; ³Good et al., 2002; ⁴Dunn and Dunn, 2007; ⁵Wechsler, 2011; ⁶Wagner et al., 1999.

α represents the Cronbach's alpha and split-half represents the Spearman–Brown prophecy formula. Reliability coefficient values above 0.71 are considered acceptable (George and Mallery, 2003). Publisher reliability information was obtained from the technical manuals and reports released by the respective companies.

we do not believe this to be a significant issue overall. The ability to record and re-watch sessions while scoring was critical given these challenges unique to the home setting.

Finally, some children felt fatigued during sessions scheduled after the child had just spent several hours on the computer during remote learning. Testers offered breaks and/or ended the session based on their judgment of the child's fatigue and engagement.

Scoring

To ensure validity, each assessment was double-scored by another tester. The second scorer watched session recordings (stored and accessed on a secure server) to verify the original scores provided by testers. If there were discrepancies between first and second scores, a core research team member who is an experienced clinician made the final scoring decision.

Scorers used an online spreadsheet to document the scoring process: the team would notate who second scored a test, their calculations of scores, any scoring discrepancies that were resolved, and any validity issues within a testing session. The scoring spreadsheet also contained formulas to automatically calculate raw scores to make the process more efficient. The

second scorer documented the final scores in REDCap. Scorers were encouraged to consult and communicate with the team whenever scoring questions or concerns arose.

Reliability

We computed Cronbach's alpha and split-half reliability for all of the standardized tasks administered in our study, except for one task where item-level information was not available from the publisher's website (DIBELS). **Table 4** provides reliability coefficients for the current study and, for comparison, the coefficients provided from the publisher for each of the subtests. The reliability coefficients for the online administration of the subtests were comparable to those reported by the publishers and are considered to be within the acceptable-good range.

Measurement Error

To further evaluate whether online administration of assessments introduced a measurement error, we calculated pairwise correlations among the standardized measures used in this study that overlapped with those administered for a different pre-pandemic in-person study in the lab (Lab Study A; **Table 5**). The comparison study (Lab Study A, Ozernov-Palchik et al., 2017) included 158 rising third-grade students with complete data for

the relevant tests. Participants for this study were recruited from 21 schools in New England and represented a demographically similar sample to that of the current study (Figure 1). The correlation patterns among the variables in both studies were similar, suggesting that the same constructs were evaluated in the online version of the assessments as in the in-person version.

Takeaways

The training process for administration and scoring of online assessments was more labor intensive than in-person studies, as there was an additional layer of developing tester competency with managing Zoom, engaging the child, and recording scores in an accessible way. Difficulties included connectivity issues and controlling for the environment (i.e., background noise, distraction). The lack of control over the child's home environment posed some reliability and validity concerns, but the flexibility of online administration also allowed for a greater ability to adapt to children's and families' individual needs. Some children may have benefited from testing in their home environment, as testing in an unfamiliar location can lead to anxiety or stress.

For those planning to implement an online testing battery in an intervention study, we recommend setting up clear and detailed systems for documentation. The amount of digital documentation was greatly increased through adaptation for virtual administration. Materials and data should be organized in the most centralized and streamlined way possible to avoid confusion and misplacement of files. Not all stimuli and record forms can be easily adapted for online administration, and alternative methods (e.g., scanning the original form) may need

to be considered depending on the assessment and availability of technology to testers and participants. It can be helpful to compile a document outlining each test, how it is administered, and links to any websites or documents needed for administration so the team has a centralized procedural document to follow.

We also recommend that before starting a new online study, researchers outline guidelines for addressing technical or environmental issues that inevitably arise (e.g., what to do if you are having trouble discerning the child's answer *via* Zoom). Technical and environmental factors cannot be eliminated when assessments are being administered virtually, but clear procedural guidance and detailed documentation during testing (e.g., noting the child's behavior and any technical issues) can help reduce reliability and validity concerns. Having an online, real-time messaging system (e.g., Slack) is also an essential tool to ensure the team is able to communicate questions and concerns. Overall, the results for the standardized measures in the current study suggest equivalent effects of online testing to those of in-person testing, which are encouraging for the potential for future online intervention studies.

INTERVENTION

Developmental researchers transitioning to a fully online intervention study will need to carefully consider how to adapt materials, train the research team (particularly if they are not located in the same place), and address difficulties that may be more likely to arise in online settings. In particular, intensive implementation of an intervention during the pandemic introduced new challenges related to privacy and disclosure. Finally, qualitative data on individuals' experiences participating in the study is important for identifying potential confounds and limitations, as well as for considering future scalability. We conclude this section by providing examples of feedback received from children and caregivers in our study.

Curriculum Adaptation

In our study, the Scaffolding Group received biweekly scaffolding sessions led by learning facilitators. For these sessions, we adapted an existing curriculum targeting oral language skills in elementary school children developed by the Language and Reading Research Consortium (LAARC; Jiang and Logan, 2019). We added verbatim scripts for the learning facilitators to read to the children. Before each session, the learning facilitators adapted these scripts to the particular text they were working on with their child. The online format allowed learning facilitators to more easily follow a script than during face-to-face communication, thereby assuring greater fidelity of implementation. We also adapted materials that are designed for use by teachers in a physical classroom to online administration. For example, we used the whiteboard feature in Zoom to draw and write words during the lesson. As part of their preparation for each lesson, the learning facilitators prepared slides with pictures of vocabulary words from the books. We embedded explicit instructions on how and when to utilize these virtual materials for each strategy. To avoid boredom and distraction, we incorporated activities to

TABLE 5 | Pairwise correlations among a sample of six variables from the current study and a comparable pre-pandemic in-person study from the same research lab.

Previous in-person sample of 3rd graders $N = 158$

| | PPVT | CELF | KBIT | Blending words | Memory for digits |
|---------------------|---------|---------|---------|----------------|-------------------|
| CELF | 0.56*** | | | | |
| KBIT | 0.30*** | 0.23** | | | |
| Blending words | 0.43*** | 0.39*** | 0.14 | | |
| Memory for digits | 0.46*** | 0.35*** | 0.27*** | 0.46*** | |
| Non-word repetition | 0.56*** | 0.35*** | 0.15 | 0.57*** | 0.51*** |

Current sample

| | PPVT | CELF | KBIT | Blending words | Memory for digits |
|---------------------|---------|---------|--------|----------------|-------------------|
| CELF | 0.43*** | | | | |
| KBIT | 0.49*** | 0.33*** | | | |
| Blending words | 0.40*** | 0.28*** | 0.22** | | |
| Memory for digits | 0.40*** | 0.23** | 0.20** | 0.29*** | |
| Non-word repetition | 0.37*** | 0.31*** | 0.24** | 0.40*** | 0.37*** |

*** $p < 0.001$, ** $p < 0.01$.

optimize child engagement during each lesson. All scaffolding sessions were recorded and stored on a secure server.

Learning Facilitator Training

We recruited and trained over 20 undergraduate students for the learning facilitator role during the study period. Students were interviewed and selected based on their experience and/or willingness to work with children and families, availability to meet consistently twice a week with their assigned participants *via* Zoom, and enthusiasm for the research. Our research team was ethnically and racially diverse, and many students were fluent in languages in addition to English. While over the summer most of the undergraduate students had full-time roles on the project, once the school year began, they had to juggle their work with their own courses and other responsibilities. Given the pandemic, many of the undergraduate students were not living on campus and completed their work from their own homes across the United States and in other countries. Our study team included many first-year students, students working in a lab for the first time, and students who did not come from a science background.

Learning facilitators underwent extensive training before being matched with participants to ensure implementation fidelity. First, we provided training in human subjects research, general strategies for working with young students, and background literature on language/reading and summer interventions. Learning facilitators were also trained on the Learning Ally audiobook platform, and began reading the books used in our study. Because all of these training sessions were remote, learning facilitators could refer back to the recordings as needed. Next, we reviewed the scripts for each lesson with learning facilitators in group meetings. Learning facilitators paired up to practice each component of the lesson with each other (e.g., check in, vocabulary instruction, and scaffolding instruction). Each learning facilitator then recorded a full practice session which was reviewed by a member of the core research team. Learning facilitators received feedback on their recorded session, and those that required additional practice were asked to record new verification videos that implemented this feedback before being assigned participants. Undergraduate students who joined our team after the first summer were matched up with an experienced learning facilitator who served as a mentor and practice partner during training and beyond.

Crucially, training did not cease when learning facilitators began working with participants. All learning facilitators attended weekly meetings where they discussed their participants' progress and troubleshooted any issues. These issues ranged from how to properly implement specific strategies in the scaffolding curriculum, to how to communicate effectively with caregivers about scheduling, to how to respond to a child that shares difficult personal circumstances (see "Child Disclosure" section, below). Learning facilitators were encouraged to reach out to members of the research team any time they wanted to review a session and discuss strategies for working with a specific child, which was facilitated by the online nature of the study. A member of the research team also spot-checked session videos and provided feedback to learning facilitators as needed to ensure intervention fidelity. Finally, we cultivated an active community

in a Slack channel, which allowed learning facilitators to post and answer questions promptly. This multi-tiered network of support enabled our team of undergraduates to thrive in the remote research setting. Notably, in addition to all of their responsibilities as learning facilitators, undergraduates also filled numerous other roles on the project such as developing proximal assessment materials, transcribing language samples, communicating with caregivers, and assisting with data maintenance.

Online Intervention

Technical Challenges During Scaffolding Sessions

The biweekly scaffolding sessions over Zoom introduced challenges unique to the virtual setting. First, researchers were dependent on the capabilities of their own and the participant's internet connection and thus had to flexibly adapt when the connection was impaired. Many participants occasionally could not see or hear their learning facilitator during crucial parts of the session, or the learning facilitator could not discern what the participant was saying from the lagging audio. Learning facilitators took many steps to troubleshoot these issues while staying on Zoom. Turning cameras off, relocating closer to the Wi-Fi router, asking for a school-provided hotspot, and even using FaceTime or phone calls in tandem with Zoom helped mediate these issues. In a few cases, learning facilitators sent the session's materials to families ahead of time to print out or download so the child would not have to wait for webpages or screen-sharing to load. Learning facilitators also supported participants who had difficulty logging into or using the Learning Ally audiobook app by asking participants to share their screens and walking them through the setup.

The online setting also enabled children to multitask during sessions. For example, there were numerous instances of participants attending sessions while siblings played video games in the same room, while friends were over, or while simultaneously doing something else on the computer. To address these distractions, learning facilitators would ask, "Are you distracted right now? How can we fix that?" and have the child come up with potential solutions. These solutions included putting on headphones, moving to another room, or asking the people around them to quiet down.

Child Disclosure

Disclosure of sensitive information occasionally came up during the testing and scaffolding sessions. In some cases this was prompted, as our study included parent and child questionnaires about experiences during the COVID-19 pandemic, negative feelings, and anxiety/depression. For example, a child disclosed that they thought about death "all the time" in response to a questionnaire item. We also anticipated that some scores on child self-report and parent-report anxiety/depression measures might fall in the clinically elevated range. In other cases, unprompted sensitive information was shared with researchers. For example, one child, when asked to use the vocabulary word 'evasive' in a sentence, said that they "used evasive action to avoid their mother hitting them." To address these expected and unexpected issues, we developed a detailed protocol for the research team to follow, overseen by a clinical psychologist who is a member of

the research team. The psychologist checked the questionnaire data for red-flag indicators (supplemental protocol⁵) weekly. If there were indicators that met our criteria for concern (e.g., anxiety or depression scores that were in the clinically elevated range), she reviewed the pertinent data available and contacted the parents/guardians to alert them about the areas of concern and potentially suggest that they consider seeking a professional consultation for further guidance, if they had not already done so. In most cases, the parents/guardians were aware that their child was struggling emotionally (and many had already sought professional help or were in the process of doing so).

If a child indicated negative thoughts or feelings directly to a research team member during a session, the research team members were instructed to notify the psychologist immediately following the session. The psychologist would then follow up with the parents/guardians as necessary. We handled the incidence when a child came up with an example sentence about trying to avoid being hit by their parent differently. Although the role of researchers in mandatory reporting is debated, many states mandate researchers working with children to report suspicion of child abuse (Allen, 2009). Consequently, we called State Child and Family Services, where the family lives, and did an anonymous screening. Based on the information we provided, we were told that “it doesn’t rise to the level of report.” We continued to monitor the child, but nothing alarming came up during the subsequent sessions.

We learned from this study that particularly when frequently working with children directly in their homes or when collecting sensitive information, issues related to children’s safety and wellbeing are likely to come up. We were fortunate to have a trained psychologist on our team who helped us develop a detailed protocol for dealing with these issues and who was responsible for communicating this information to families in a non-alarming but informative manner. Although not always mandated by the IRB, every study that involves children should include detailed procedures for handling sensitive information. Additionally, particularly for online studies that span several states, it is important to know which agency handles suspicions of potential abuse or neglect and what responsibilities researchers working with children have in that state.

Finally, it is important to support team members who may hear from children about difficult challenges they are facing. Most research assistants do not have mental health training, and thus may experience stress or other reactions to instances of child disclosure. Our learning facilitators were undergraduate students who themselves had been dealing with unprecedented challenges related to the pandemic. We addressed these potential challenges explicitly during training and through encouraging continuous communication within the team throughout the study, and by clearly indicating who to contact if such an issue arose. On our Slack channel and during weekly meetings, team members shared their experiences, debriefed, and coached each other on how to best respond to participants. In specific instances (described below), the clinical psychologist on our team provided one-on-one support to team members.

⁵<https://osf.io/6urmx/>

Qualitative Caregiver and Child Experiences

Child Reflections

At the end of the 8 weeks of meeting with learning facilitators, many participants did not want the study to end. When one learning facilitator started the last session with her student by saying, “Are you ready for our last lesson today?” the participant responded, “Yes, but I don’t want it to be our last lesson,” and ended up signing off the call by saying, “Okay, love you, see you, bye!” Another participant who always brought his favorite stuffed animal, Teddy, to the sessions remarked that, “Teddy is sad,” when saying their goodbyes at their final meeting.

Many children reported enjoying the study experience, even if they did not enjoy their regular school-related activities or reading. During her final session, one student remarked “I hate school! School is evil.” The learning facilitator said “Well, this is like school and this was really fun!” to which the participant said, “This wasn’t evil.” One participant who had previously stated he did not enjoy reading told his parent at the 7-week mark: “You know what’s so great about the audiobooks mum? It’s that they’re able to go into such more details than movies!” The parent expanded on this: “I cannot express the joy it brings me to hear my son starting conversations with me about stories he’s read. Last week he wanted to recount some various storylines to me from books. To [say] that we’ve been enjoying the experience is an understatement. Thank you.”

Many children also faced pandemic-related challenges that affected them during the course of the study. In addition to being out of school and having their social lives change, a few had family members who were directly affected by the virus. For instance, one participant was living with an uncle who had COVID-19. During one session, she told her learning facilitator, “People are in my house and it’s difficult for me and my mom because, you know, my uncle is going to die. They want to help him, but they can’t.” One week later, during the routine check-in, the learning facilitator asked how she was doing and the participant said she was sad; “Yesterday, my uncle died. We saw him and, like, it’s sad for me since I [have known] him since I was a kid. Me and my mom [were] crying.” Her learning facilitator expressed her condolences, letting the child know that this is an extremely difficult time. She made sure to offer the participant an opportunity for breaks, instating a codeword of “rainbow sunshine.” The learning facilitators adapted to meet the participants where they were at emotionally and mentally each session, knowing that the pandemic affected everyone’s lives differently, and were generally a welcoming, consistent presence in the participants’ lives for the duration of the study. Importantly, children participating in our research always come into our sessions with a variety of experiences. While the pandemic led to more consistent challenges among our participants, these difficult experiences – death, illness, stress, financial insecurity – should always be on the research team’s radar. At the end of the study, participants in the Scaffolding Group reported generally positive experiences (Table 6).

Caregiver Reflections

At the end of the study, caregivers filled out a reflection survey about their experience in the study. In general, caregivers of children in the Scaffolding Group did not find it difficult for their child to have biweekly online meetings with their learning facilitator (Table 7).

Caregivers in both the Scaffolding Group and Audiobooks-only group likewise provided open-ended responses about their experiences in the study. Selected representative responses are included below (Table 8). Participants in the Audiobooks-only condition did not meet regularly with a learning facilitator, but they did receive weekly messages with updates on reading milestones and suggested book titles to read.

As reflected in these responses, caregivers in both groups had many positive experiences in the study. The remote learning environment fostered feelings of social isolation and loneliness for many children (as reflected in our surveys). In the Scaffolding Group, caregivers generally commented on interactions with the learning facilitators, and suggested that the connections forged between children and learning facilitators in our study may have helped ameliorate some of the negative socio-emotional consequences of the pandemic. This positive feedback is useful as we consider implementing future online interventions. In the Audiobooks-only Group, positive feedback focused on the reading experience and book selection.

Challenges were modest for both groups, and some challenges were not unique to the remote nature of the study. For instance, caregivers of children in the Scaffolding Group reported some difficulty finding time for sessions and getting their child to read the books, and some caregivers commented on the challenging nature of the vocabulary. In the Audiobooks-only Group, some caregivers noted that their child was not always interested

TABLE 8 | Caregiver experiences in scaffolding and audiobooks-only groups.

| | Scaffolding group | Audiobooks-only group |
|--|---|---|
| What did your child enjoy most in this study? | <p>"My child enjoyed all aspects of the study. He is proud to tell others that he is participating in a study. He is very excited to be paid by gift certificates. He loves how he can access any book of his choosing. He enjoyed the experience of meeting weekly and discussing the books with someone."</p> <p>"My son really enjoyed meeting with the learning facilitator and was sad to learn he would not be meeting with the facilitator anymore. He loved the books and the platform though I was hoping he would read more without me reminding him."</p> <p>"He enjoyed being introduced to books he may not have otherwise picked out to read. He also liked meeting with his facilitator. He is a social kid and the pandemic has been hard, so seeing [his learning facilitator] was a highlight of the week."</p> | <p>"He definitely enjoyed listening to the books that were recommended the best!!"</p> <p>"It allowed her to be independent with her nightly reading."</p> <p>"She enjoyed engaging with the tester. She enjoyed being able to pick her own book and listen on her own. This contributed to family conversations regarding the stories she listened too."</p> <p>"He really enjoyed the interviews and listening to/reading along w/Learning Ally. I would like to continue it. He would often have siblings gathered around, reading too."</p> <p>"Es una experiencia bonita para los niños, por que es una manera de leer sin leer o sea escuchando, es diferente pero me gusta, hasta la niña de segundo grado quería escuchar los libros, me gusto mucho, gracias sigan así ayudando a niños a que le den importancia a la lectura."</p> <p><i>Translation: "It is a beautiful experience for the kids because this way they can read with listening, it's different but I like it. Even my second grade daughter wanted to listen to the books. I enjoyed it a lot. Keep up the good work"</i></p> |
| What did your child find most challenging in this study? | <p>"She found the questions and vocabulary hard."</p> <p>"He is not used to listening to books and using the app required more setup time since he had to use his laptop, so it was something we had to remind him to do."</p> <p>"Finding time to read the books, especially without distraction"</p> <p>"Twice weekly meetings with the facilitator was a lot for our schedule"</p> <p>"She sometimes did not want to stop what she was doing to attend scaffolding. Also wanted to socialize and share other things with Facilitator not fully focused on session"</p> | <p>"She did not like listening to books she had no interest in."</p> <p>"Trying to read/listen to the books she was not immediately interested in. I challenged her to try at least half of the book to see if it improved and she did not like that."</p> <p>"The second book didn't hold her interest"</p> <p>"Mostly technical problems"</p> <p>"Por las circunstancias pasa mucho tiempo conectado a algún dispositivo electrónico y aveces solo quería hacer otra cosa, en circunstancias normales creo sería su actividad favorita."</p> <p><i>Translation: "Because of the circumstances he spent a lot of time connected to an electronic device and sometimes he wanted to do something else. Under normal circumstances this might have been his favorite activity"</i></p> |

TABLE 6 | Child experiences in scaffolding group.

| How much did you like meeting with your learning facilitator? | |
|---|------------|
| Not at all | 1 (1.8%) |
| A little bit | 3 (5.3%) |
| Sometimes | 11 (19.3%) |
| A lot | 42 (73.7%) |
| How often did you feel like you learned new words with your learning facilitator? | |
| Not at all | 1 (1.8%) |
| A little bit | 5 (8.8%) |
| Sometimes | 11 (19.3%) |
| A lot | 40 (70.2%) |

TABLE 7 | Caregiver experiences in scaffolding group.

| Was it challenging to get your child to meet with their learning facilitator? | |
|---|------------|
| Not at all | 50 (80.6%) |
| A little bit | 9 (14.5%) |
| Sometimes | 3 (4.8%) |
| A lot | 0 |

in the recommended books. This group received the same book recommendations as the Scaffolding Group, but they did not discuss the books with a learning facilitator, which we hypothesized would impact their engagement. The Audiobooks-only Group also received only weekly updates; thus, they were unable to change books that did not interest them as easily as participants in the Scaffolding Group. Some caregivers also reported technical difficulties during and after the study. We relayed all technical issues to the audiobook company, and they worked with us and the caregivers to find solutions.

Takeaways

To properly measure intervention effects, we needed to ensure that both participants and learning facilitators were properly supported for an online intervention. Particularly for our learning facilitators, who had no previous experience implementing interventions, extensive training and open communication with supervisors and peers was critical. We found that weekly meetings and an internal study Slack channel provided opportunities for learning facilitators to learn from one another and troubleshoot issues. Consistent communication and chances to check-in were crucial since we could not share a physical lab space. Video recording of all sessions allowed for ensuring fidelity of implementation and consistency across different learning facilitators and sessions.

The Scaffolding Group provided useful lessons for other researchers conducting studies with frequent online meetings. Researchers should expect some sessions to have distractions and technical difficulties; thus, it is important to have plans in place to ensure the fidelity of the study. Families reported only modest difficulties with study demands, and feedback from caregivers and children were overall positive. Indeed, many children felt comfortable sharing even highly personal information with their learning facilitators. Researchers should establish clear protocols for how to deal with sensitive information shared by children and families, particularly for studies that involve lots of online interactions.

DISCUSSION

We implemented a fully remote RCT intervention (final $N = 255$ third and fourth graders, ages 8–10 years) targeting children's language comprehension skills, which we described as a case study to explore various factors involved in conducting an online intervention study. We have summarized the challenges we faced, solutions we devised, and considerations for future research. Although our project represents a specific case study, and the implications should be considered carefully, we believe that the unique context of our study, its intensity and scale, and our diverse recruitment efforts allow us to derive 'lessons learned' that could be useful for others embarking on a similar project. We conclude by discussing what we believe to be the three main tradeoffs to think about when deciding whether and how to implement an online intervention study with a developmental sample (Figure 5).

Internal vs. External Validity

An important goal of RCTs is to design and evaluate carefully controlled interventions that allow researchers to understand the precise causal mechanisms by which an intervention leads to learning gains. However, this can come at a cost – sometimes, the more controlled the intervention, the less likely it is to work in the “real world.” As with any other type of study, an online RCT intervention requires researchers to consider tradeoffs between internal validity (how well the experiment tests what it is meant to test and is not influenced by other factors) and external validity (how well the experiment replicates in a natural environment).

Most developmental studies optimize internal validity by conducting studies in labs. These studies are well-poised to isolate the precise mechanism or phenomenon researchers are interested in studying. However, there are also drawbacks to in-lab studies that are particularly relevant for researchers interested in conducting RCTs. In-lab developmental studies typically rely on convenience samples, which tend to be homogenous, thereby limiting generalization to other populations (Bornstein et al., 2013). Furthermore, due to multiple practical considerations (e.g., space limitations, transportation, scheduling issues), in-person studies tend to have smaller sample sizes than what is possible in online data collection. Finally, the ecological validity of such studies has been criticized – and the implications for what developmental processes look like in messy and unpredictable real-world settings, such as learning in a child's home, are limited (Lortie-Forgues and Inglis, 2019). Thus, while implementing an RCT study online in children's homes requires giving up some of the control of in-lab experiments and introduces additional noise, the tradeoff is that these studies can be more naturalistic and lead to increased sample diversity.

Especially important to consider for intervention studies is generalizability of effectiveness. On the other side of the spectrum from carefully controlled in-lab studies are large-scale educational RCT studies that implement interventions in schools and childcare settings. These studies tend to have higher external validity, but a side effect is increased noise. These studies often build on pilot studies that establish the value of a particular intervention under tightly controlled conditions, but they tend to have small efficacy in these real-world settings (Lortie-Forgues and Inglis, 2019). There are many reasons for this. For example, school settings may be prohibitive of careful sample selection using stringent exclusion criteria (i.e., one child in a classroom receives the intervention while another child does not). Although there are design and statistical methods to overcome these issues (e.g., Regression Discontinuity Design; Lee and Munk, 2008), online intervention studies can bypass them altogether by working with eligible children in their own homes, which expands the pool of participants who are eligible to participate while also allowing the use of specific eligibility criteria and random group assignments. Similarly, it is more difficult to monitor and ensure implementation fidelity of programs when working in complex formal institutional environments such as schools, as compared to negotiating logistics with a child-researcher duo. In our study, we were able to overcome these obstacles because we could closely monitor research activities *via* direct and continuous communication and video recording,



FIGURE 5 | Tradeoffs for online intervention studies with developmental populations.

and to document possible threats to validity during the various aspects of the study (e.g., background noise, child distraction, connectivity issues, implementation fidelity, etc.).

Thus, we suggest that the online implementation of intervention studies could improve the internal validity of such studies while maintaining their external validity. In online studies, the research team can operate within a well-controlled lab environment, while working with participants in natural, ecologically valid settings. We discussed several potential threats to the validity of our study, such as background noise and technological challenges that could impact reliable data collection. Based on the comparison of the reliability scores for the current study and in-lab studies, however, online data collection resulted in equally reliable data collection, supporting the feasibility of maintaining internal validity in remote developmental research. The increased racial and socioeconomic diversity of the current sample, as compared to in-lab samples, suggests that we were able to achieve greater ecological validity. Furthermore, our study was conducted entirely in children's natural context – in their own homes – supporting its potential efficacy in real-world settings.

Available Research Resources vs. Participant Engagement

Implementing an RCT can be resource intensive – e.g., researchers' time, project budget, number of personnel – and often requires making decisions regarding how many resources to devote in order to maximize participant engagement and retention. Participant engagement can be measured across different levels (Matthews et al., 2011). Recruitment is one such measure that considers the reach of the study to the target population. Many educational intervention studies rely on school partnerships for recruitment, which can be an effective strategy for recruiting a large number of children from diverse educational environments. However, establishing school partnerships requires substantial time and energy. The research team first has to clearly communicate the goals of the intervention and the benefits to that school's community in order to get buy-in from school leaders and educators. This process typically relies on existing relationships with schools and institutional familiarity, which might be more difficult for a new investigator to establish. Even when schools are interested

in a potential partnership, the bureaucratic processes can be extensive before the study can get started. It can also be difficult to randomly assign students to conditions within a school because once a school is enthusiastic about an intervention, the school often wants all their students to be placed in the intervention condition.

On the other hand, many developmental science studies recruit participants directly through advertisements and social media (Hurwitz et al., 2017). Social media recruitment efforts can reach a wide pool of potential participants at a reasonably low cost. Our social media reach was extensive, reaching people from hundreds of different zip codes across the United States, but this required intentional targeted advertising. Based on our recruitment data, through school partnerships and social media, we successfully reached the participant demographic we set to recruit.

Enrollment, retention, and intervention adherence are additional types of engagement, each with its own set of challenges. Our enrollment and retention outcomes were less successful than our recruitment reach. Our final sample, although still very diverse, was not representative of the diversity in schools and communities we targeted in our recruitment. For example, household income eligibility for free/reduced lunch is around \$52,000. Although we targeted schools and communities with a high proportion of free/reduced lunch eligibility, we ended up with a median income with the \$80,000–120,000 range. Thus, even though we allocated almost all of our recruiting budget and efforts to recruit lower-SES participants, our final enrollment was not skewed toward this demographic. Retention and intervention adherence represent two of the most critical factors to ensure the validity of intervention studies (Slack and Draugalis, 2001) and are most difficult to achieve when working with disadvantaged communities. Ensuring participant engagement in such communities is resource-intensive, requiring a substantial recruitment budget, a large and well-trained research team, and attractive incentives for participation.

There is a large body of evidence from parenting programs targeting underserved communities that show how program-level factors (e.g., team member composition, level of family support provided) interact with participant factors (e.g., SES, job demands, perception of research, language barriers) in ensuring enrollment and retention (Whittaker and Cowley,

2012; Hackworth et al., 2018). Families, especially those from lower-SES backgrounds, are more likely to enroll and stay in a program, for example, if they have an experienced research liaison who supports them in identifying and overcoming barriers to participation (Rivas-Drake et al., 2016; Hackworth et al., 2018). Our full-time, bilingual coordinators were available to check in and assist families using preferred communication methods, and researchers assisted families with troubleshooting the apps for the intervention. Clear communication on research objectives and the theoretical foundation of the intervention is important for reducing perceptual barriers to participation (Barlow et al., 2003; Moran et al., 2004). Professionalism and experience of team members (Hackworth et al., 2018), as well as their representativeness of the target community (Gray, 2002), were additional factors that ensured engagement. During our consent process, as well throughout the study, researchers were available to answer questions. We also hosted several information sessions for teachers and administrators in our partner district, as well as a bilingual (Spanish/English) session for parents at one of our partner schools. Intervention effects have been more significant in well-resourced studies, as compared to studies with fewer resources (Kim and Quinn, 2013). In general, across studies, there is an agreement that intervention programs targeting lower-SES communities require careful considerations of various factors that could affect direction of resources toward alleviating these barriers.

Online research may seem like a low-resource opportunity for obtaining larger, more diverse samples. With the advent of online platforms for developmental studies (e.g., Discoveries Online; Lookit), unmoderated research studies have become increasingly popular. Such studies, which allow participants to complete tasks on their own time and without the researcher's direct involvement, front-load their resources for design but require minimal resources for implementation. We caution, however, that families from underrepresented backgrounds may still face greater barriers to engaging in such studies than participants that are typically included in research studies, and we echo calls to actively work toward providing support and internet access for these populations (Lourenco and Tasimi, 2020; Sheskin et al., 2020). This is particularly pertinent for longitudinal and intervention studies that require substantial researcher moderation in order to be successful. Indeed, a similar online intervention during the pandemic that did not explicitly target a diverse sample based on SES ended up with almost all mothers with at least a 4-year college degree (Bambha and Casasola, 2021). We found that even children in school systems that did provide devices and internet access sometimes experienced technical difficulties in our study. Thus, while online RCTs can remove certain resource constraints (such as space and travel compensation), researchers should expect to invest significant time and effort to achieve diverse samples and ensure their participation.

Geographic Diversity vs. Digital Divide

Online study participation with children, although not always feasible, can significantly increase sample diversity by allowing easy access regardless of a family's geographic location and

by minimizing caregivers' time commitment (Rhodes et al., 2020; Sheskin et al., 2020). This is particularly crucial for longitudinal studies that include multiple sessions and a significant time commitment. Online developmental studies have recruited more diverse samples than in-lab developmental studies (e.g., Scott and Schulz, 2017; Scott et al., 2017), including more geographically diverse samples (Bambha and Casasola, 2021). Our study recruited participants from 26 different states in the United States (Figure 2), and our sample was comparable to or better than our prior in-lab studies in terms of socioeconomic and racial diversity (Figure 1). However, the accessibility of online study participation is still challenging for many families (Lourenco and Tasimi, 2020). Prior to the start of the pandemic, almost a third of public K-12 students in the United States lacked adequate internet access and/or an adequate device for distance learning (Chandra et al., 2020). While some school systems provided children with computers and internet access to enable remote learning, many children still lack technology that would enable them to participate in an online intervention study. We unfortunately had to exclude interested families who lacked a computer or tablet at home due to our assessment battery. Furthermore, the "digital divide" – that is, the gap between people who have computer and internet access and those who do not – is not equally distributed across geographic boundaries and demographic groups (Van Dijk, 2020). 37% of students in rural communities in the United States lack adequate internet connectivity at home, compared to 21% of students in urban environments (Chandra et al., 2020). Many of our participants struggled with internet connectivity issues and other technological challenges over the course of the study. Thus, it is important to take into account not only whether participants have access, but also whether they have complete access to these studies. In contrast, intervention studies that do not require the family to learn about the study and participate through their own technological platforms (such as most in-school interventions) allow researchers to ensure all participants in a constrained location can participate. Yet in-person interventions are not equally accessible to all geographic regions either – most of these studies take place near research institutions. One solution is to provide participants with the technology they need to participate in online research studies (Lourenco and Tasimi, 2020). Though adding additional costs to the study budget, providing devices with mobile data may lead to more representative samples as well as better data quality. For example, several large-scale projects have successfully deployed mobile devices loaded with educational content in rural locations in the United States and around the world, like small villages in Africa (Breazeal et al., 2016; Uchidiuno et al., 2018). This tradeoff may be worth the cost, particularly for home-based intervention studies. Online studies allow for geographic diversity of the research team as well. Our study team worked from multiple time zones, which allowed us to accommodate participants from across the United States. This also opens up the possibility for recruiting community members to be part of the research team. This type of participatory research may lead to

higher recruitment, retention, and validity of intervention studies (Levac et al., 2019).

CONCLUSION

In response to the COVID-19 pandemic we conducted a scalable online RCT intervention study with children from diverse backgrounds across the United States. In this paper, we summarized the challenges we encountered and the tradeoffs to consider when implementing such studies. Despite possible threats to the internal validity of our study, difficulties in reaching demographically diverse populations, and resource-exhaustive efforts to support participant engagement and retention, we were able to conduct a study that provided educational support during a challenging time for both children and their caregivers. With the aforementioned considerations and tradeoffs in mind, we believe that fully remote intervention studies are a worthwhile endeavor for developmental researchers, and we expect to see more of them in the future.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repository and accession number(s) can be found below: OSF: <https://osf.io/zac9d/>.

ETHICS STATEMENT

Studies involving human participants were reviewed and approved by the Committee on the Use of Human Experimental Subjects at MIT. Informed consent to participate in this study was provided by the participants' parents or legal guardians.

REFERENCES

- Allen, B. (2009). Are researchers ethically obligated to report suspected child maltreatment? a critical analysis of opposing perspectives. *Ethic. Behav.* 19, 15–24. doi: 10.1080/10508420802623641
- Bambha, V. P., and Casasola, M. (2021). From Lab to Zoom: Adapting training study methodologies to remote conditions. *Front. Psychol.* 2021:12. doi: 10.3389/fpsyg.2021.694728
- Barlow, J., Coren, E., and Stewart-Brown, S. (2003). Parent-training programmes for improving maternal psychosocial health. *Coch. Datab. Syst. Rev.* 2003:4.
- Barton, E. E., and Fetting, A. (2013). Parent-implemented interventions for young children with disabilities: a review of fidelity features. *J. Early Interv.* 35, 194–219.
- Bornstein, M. H., Jager, J., and Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Dev. Rev.* 33, 357–370. doi: 10.1016/j.dr.2013.08.003
- Breazeal, C., Morris, R., Gottwald, S., Galyean, T., and Wolf, M. (2016). Mobile devices for early literacy intervention and research with global reach. *ACM Conf.* 2016, 11–20. doi: 10.1002/cad.20225
- Casler, K., Bickel, L., and Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comp. Hum. Behav.* 29, 2156–2160.
- Chandra, S., Chang, A., Day, L., Fazlullah, A., Liu, J., McBride, L., et al. (2020). *Closing the K–12 digital divide in the age of distance learning*. Boston, MA: Common Sense and Boston Consulting Group.

AUTHOR CONTRIBUTIONS

OO-P and HO contributed equally to the study idea, design, implementation, and manuscript preparation. XA and HK assisted with study design. XA, HK, JS-F, KW, YCT, NG, and JD assisted with study implementation, and each contributed to writing one or more sections and designed figures for this manuscript. JG contributed to the study idea and design, and provided feedback on this manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Chan Zuckerberg Initiative for the Reach Every Reader project (<https://www.gse.harvard.edu/reach-every-reader>), the National Science Foundation (Graduate Research Fellowship 1745302 to HO), and the National Institutes of Health (F32-HD100064 to OO-P).

ACKNOWLEDGMENTS

We are grateful for the partnership with Learning Ally, for helpful suggestions from James Kim and Tiffany Hogan. We would also like to thank our undergraduate research assistants: Tolu Asade, Cherry Wang, Sophia Angus, Bhuvna Murthy, Maycee McClure, Sehry Khan, Hilary Zen, Sarah Abodalo, Elizabeth Carbonell, Shruti Das, Erika Leasher, Yoon Lim, Emmi Mills, Zoë Elizee, Camille Uldry, Shelby Laitipaya, Alexis Cho, Gabriella Aponte, Harley Yoder, Dana Osei, Niki Kim, and Joy Bhattacharya; our testers: Amanda Miller, David Bates, Ross Weissman, Joohee Baik, June Okada, William Oliver, and Harriet Richards; and our additional team members: Isaac Treves, Cindy Li, Kristen Wehara, Brooke Goldstein, and Ada Huang. Finally, we are grateful to the families for their time and participation.

- Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., et al. (2021). Moderated online data-collection for developmental research: methods and replications. *Front. Psychol.* 12:734398. doi: 10.3389/fpsyg.2021.734398
- Dunn, L. M., and Dunn, D. M. (2007). *Peabody picture vocabulary test-fourth edition (PPVT-4)*. Circle Pines, MN: AGS.
- Fixen, D., Naom, S., Blase, K., Friedman, R., and Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature*. Tampa: University of South Florida The National Implementation Research Network.
- George, F., and Mallery, M. (2003). *Quantitative method for estimating the reliability of data*. Kalamazoo, MI: Western Michigan University.
- Gillen, N. A., Siow, S., Lepadatu, I., Sucevic, J., Plunkett, K., and Duta, M. (2021). Tapping into the potential of remote developmental research: introducing the OxfordBabylab app. *PsyArXiv* 2021:1. doi: 10.1002/9780470773307.ch1
- Good, R. H., Kaminski, R. A., Smith, S., and Laimon, D. (2002). *Dynamic indicators of basic early literacy skills: DIBELS*. Houston, TX: Dynamic Measurement Group.
- Gray, B. (2002). Emotional labour and befriending in family support and child protection in Tower Hamlets. *Child Family Soc. Work* 7, 13–22. doi: 10.1046/j.1365-2206.2002.00222.x
- Hackworth, N. J., Matthews, J., Westrupp, E. M., Nguyen, C., Phan, T., Scicluna, A., et al. (2018). What influences parental engagement in early intervention? Parent, program and community predictors of enrolment, retention and involvement. *Prev. Sci.* 19, 880–893. doi: 10.1007/s11121-018-0897-2

- Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., et al. (2019). The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* 95:103208. doi: 10.1016/j.jbi.2019.103208
- Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* 42, 377–381. doi: 10.1016/j.jbi.2008.08.010
- Heinrichs, N., Bertram, H., Kuschel, A., and Hahlweg, K. (2005). Parent recruitment and retention in a universal prevention program for child behavior and emotional problems: Barriers to research and program participation. *Prevent. Sci.* 6, 275–286. doi: 10.1007/s11121-005-0006-1
- Hurwitz, L. B., Schmitt, K. L., and Olsen, M. K. (2017). Facilitating development research: Suggestions for recruiting and re-recruiting children and families. *Front. Psychol.* 8:1525. doi: 10.3389/fpsyg.2017.01525
- Jiang, H., and Logan, J. (2019). Improving reading comprehension in the primary grades: mediated effects of a language-focused classroom intervention. *J. Speech Lang. Hear. Res.* 62, 2812–2828. doi: 10.1044/2019_JSLHR-L-19-0015
- Kaufman, A. S. (2004). *Kaufman brief intelligence test—second edition (KBIT-2)*. Circle Pines, MN: American Guidance Service.
- Kim, J. S., and Quinn, D. M. (2013). The effects of summer reading on low-income children's literacy achievement from kindergarten to grade 8: A meta-analysis of classroom and home interventions. *Rev. Educ. Res.* 83, 386–431.
- Kizilcec, R. F., Reich, J., Yeomans, M., Dann, C., Brunskill, E., Lopez, G., et al. (2020). Scaling up behavioral science interventions in online education. *Proc. Natl. Acad. Sci.* 117, 14900–14905. doi: 10.1073/pnas.1921417117
- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educ. Res.* 49, 241–253. doi: 10.3102/0013189x20912798
- Lee, H., and Munk, T. (2008). *Using regression discontinuity design for program evaluation*. Rockville, MD: Research Blvd, 3–7.
- Levac, L., Ronis, S., Cowper-Smith, Y., and Vaccarino, O. (2019). A scoping review: The utility of participatory research approaches in psychology. *J. Comm. Psychol.* 47, 1865–1892. doi: 10.1002/jcop.22231
- Lingwood, J., Levy, R., Billington, J., and Rowland, C. (2020). Barriers and solutions to participation in family-based education interventions. *Internat. J. Soc. Res. Method.* 23, 185–198. doi: 10.1080/13645579.2019.1645377
- Lorenc, T., Petticrew, M., Welch, V., and Tugwell, P. (2013). What types of interventions generate inequalities? Evidence from systematic reviews. *J. Epidemiol. Comm. Health* 67, 190–193. doi: 10.1136/jech-2012-201257
- Lortie-Forgues, H., and Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educ. Res.* 48, 158–166.
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: Conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Manz, P. H., Hughes, C., Barnabas, E., Bracaliello, C., and Ginsburg-Block, M. (2010). A descriptive review and meta-analysis of family-based emergent literacy interventions: To what extent is the research applicable to low-income, ethnic-minority or linguistically-diverse young children? *Early Childh. Res. Q.* 25, 409–431.
- Matthews, E. C., Fox, S., Hackworth, N., Kitanovski, M., and Vista, A. (2011). *The Parenting Research Centre. We wish to acknowledge the valuable contributions and support of: Iris Crook, Community Development Worker, Family & Children's Services, Yarra Ranges Shire; Georgina Devereaux, Playgroup Support and Development Officer, Frankston City Council*. Victoria: Partnerships Division Office for Children and Portfolio Coordination Department of Education and Early Childhood Development Melbourne.
- Moran, P., Ghate, D., Van Der Merwe, A., and Policy Research Bureau (2004). *What works in parenting support?: A review of the international evidence*. London: DfES Publications.
- Nicholson, L. M., Schwirian, P. M., Klein, E. G., Skybo, T., Murray-Johnson, L., Eneli, I., et al. (2011). Recruitment and retention strategies in longitudinal clinical studies with low-income populations. *Contemp. Clin. Trials* 32, 353–362. doi: 10.1016/j.cct.2011.01.007
- Nielsen, M., Haun, D., Kärtner, J., and Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *J. Exp. Child Psychol.* 162, 31–38. doi: 10.1016/j.jecp.2017.04.017
- O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Rev. Educ. Res.* 78, 33–84.
- Ozernov-Palchik, O., Norton, E. S., Sideridis, G., Beach, S. D., Wolf, M., Gabrieli, J. D., et al. (2017). Longitudinal stability of pre-reading skill profiles of kindergarten children: Implications for early screening and theories of reading. *Dev. Sci.* 20:e12471. doi: 10.1111/desc.12471
- Pollack, C., Wilmot, D., Centanni, T., Halverson, K., Frosch, I., D'Mello, A., et al. (2021). Anxiety, motivation, and competence in mathematics and reading for children with and without learning difficulties. *Front. Psychol.* 2021:704821. doi: 10.3389/fpsyg.2021.704821
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Rivas-Drake, D., Camacho, T. C., and Guillaume, C. (2016). Just good developmental science: Trust, identity, and responsibility in ethnic minority recruitment and retention. *Adv. Child Dev. Behav.* 50, 161–188. doi: 10.1016/bbs.acdb.2015.11.002
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (Part 2): Assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/opmi_a_00001
- Scott, K., and Schulz, L. (2017). Lookit (part 1): A new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/opmi_a_00002
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Slack, M. K., and Draugalis, J. R. Jr. (2001). Establishing the internal and external validity of experimental studies. *Am. J. Health-Syst. Pharm.* 58, 2173–2181. doi: 10.1093/ajhp/58.22.2173
- Uchidiuno, J., Yarzebinski, E., Madaio, M., Maheshwari, N., Koedinger, K., and Ogan, A. (2018). *Designing appropriate learning technologies for school vs home settings in tanzanian rural villages*. Seattle: ACM SIGCAS Conference, 1–11.
- Van Dijk, J. (2020). *The digital divide*. Hoboken: John Wiley & Sons.
- Veinot, T. C., Mitchell, H., and Ancker, J. S. (2018). Good intentions are not enough: How informatics interventions can worsen inequality. *J. Am. Med. Inform. Assoc.* 25, 1080–1088. doi: 10.1093/jamia/ocy052
- Wagner, R. K., Torgesen, J. K., Rashotte, C. A., and Pearson, N. A. (1999). *Comprehensive test of phonological processing: CTOPP*. Austin, TX: Pro-ed inc.
- Wechsler, D. (2011). *WASI-II: Wechsler abbreviated scale of intelligence*. London: PsychCorp.
- Whittaker, K. A., and Cowley, S. (2012). An effective programme is not enough: A review of factors associated with poor attendance and engagement with parenting support programmes. *Child. Soc.* 26, 138–149. doi: 10.1111/j.1099-0860.2010.00333.x
- Wiig, E. H., Secord, W. A., and Semel, E. (2013). *Clinical evaluation of language fundamentals: CELF-5*. San Antonio, TX: Pearson.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer JFK declared a shared affiliation, with one of the authors OO-P to the handling editor at the time of the review.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ozernov-Palchik, Olson, Arechiga, Kentala, Solorio-Fieldler, Wang, Torres, Gardino, Dieffenbach and Gabrieli. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Comparing Online Webcam- and Laboratory-Based Eye-Tracking for the Assessment of Infants' Audio-Visual Synchrony Perception

Anna Bánki^{1*}, Martina de Eccher², Lili Falschlehner¹, Stefanie Hoehl¹ and Gabriela Markova¹

¹ Department of Developmental and Educational Psychology, Faculty of Psychology, University of Vienna, Vienna, Austria,

² Department for Psychology of Language, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen, Göttingen, Germany

OPEN ACCESS

Edited by:

Sho Tsuji,
The University of Tokyo, Japan

Reviewed by:

Jessica Tan,
The University of Tokyo, Japan
Cécile Issard,
Columbia University, United States

*Correspondence:

Anna Bánki
anna.banki@univie.ac.at

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 30 June 2021

Accepted: 06 December 2021

Published: 11 January 2022

Citation:

Bánki A, de Eccher M,
Falschlehner L, Hoehl S and
Markova G (2022) Comparing Online
Webcam- and Laboratory-Based
Eye-Tracking for the Assessment
of Infants' Audio-Visual Synchrony
Perception.
Front. Psychol. 12:733933.
doi: 10.3389/fpsyg.2021.733933

Online data collection with infants raises special opportunities and challenges for developmental research. One of the most prevalent methods in infancy research is eye-tracking, which has been widely applied in laboratory settings to assess cognitive development. Technological advances now allow conducting eye-tracking online with various populations, including infants. However, the accuracy and reliability of online infant eye-tracking remain to be comprehensively evaluated. No research to date has directly compared webcam-based and in-lab eye-tracking data from infants, similarly to data from adults. The present study provides a direct comparison of in-lab and webcam-based eye-tracking data from infants who completed an identical looking time paradigm in two different settings (in the laboratory or online at home). We assessed 4-6-month-old infants ($n = 38$) in an eye-tracking task that measured the detection of audio-visual asynchrony. Webcam-based and in-lab eye-tracking data were compared on eye-tracking and video data quality, infants' viewing behavior, and experimental effects. Results revealed no differences between the in-lab and online setting in the frequency of technical issues and participant attrition rates. Video data quality was comparable between settings in terms of completeness and brightness, despite lower frame rate and resolution online. Eye-tracking data quality was higher in the laboratory than online, except in case of relative sample loss. Gaze data quantity recorded by eye-tracking was significantly lower than by video in both settings. In valid trials, eye-tracking and video data captured infants' viewing behavior uniformly, irrespective of setting. Despite the common challenges of infant eye-tracking across experimental settings, our results point toward the necessity to further improve the precision of online eye-tracking with infants. Taken together, online eye-tracking is a promising tool to assess infants' gaze behavior but requires careful data quality control. The demographic composition of both samples differed from the generic population on caregiver education: our samples comprised caregivers with higher-than-average education levels, challenging the notion that online studies will *per se* reach more diverse populations.

Keywords: infant eye-tracking, method comparison, online research, preferential looking, synchrony perception

INTRODUCTION

The current worldwide pandemic situation necessitated a change to online data collection methods for developmental psychology research (Lourenco and Tasimi, 2020). A switch to remote data collection has been particularly challenging for infant studies that mostly rely on in-person observation methods (Rhodes et al., 2020). Initiatives to move developmental science online started to increase rapidly during the last year (Leshin et al., 2020; Sheskin et al., 2020), building on existing moderated (Sheskin and Keil, 2018) and unmoderated remote research attempts and experiment platforms (Scott and Schulz, 2017; Scott et al., 2017; Semmelmann et al., 2017; Tran et al., 2017) in the field. New tools and platforms for moderated and unmoderated online studies targeting developmental populations have also recently emerged (Rhodes et al., 2020; Lo et al., 2021; Oliver and Pike, 2021; Su and Ceci, 2021).

Moderated or synchronous online research is based on researchers collecting data *via* direct interaction with participants (i.e., *via* videoconference), whereas unmoderated or asynchronous online studies do not require the presence of experimenters (Sheskin et al., 2020). Moderated and unmoderated procedures could also be combined: experimenters may instruct parents in a live video call on how to carry out the experimental task and provide support with the participant's set-up or troubleshooting technical issues (Smith-Flores et al., 2021). Moderated online studies have been successfully adapted for older children (Chuey et al., 2020; Kominsky et al., 2021b; Richardson et al., 2021; Yamamoto et al., 2021), but could be challenging to realize with infants if the study design involves social interaction (Lo et al., 2021). Prior research shows that infants only become able to initiate joint visual attention by the age of 16 months during online interactions (McClure et al., 2018), thus moderated experiments mostly rely on observations of parental and infant behavior (Libertus and Violi, 2016; Daghighi et al., 2020; Oliver and Pike, 2021). To run experimental tasks with infants online, unmoderated data collection has advantages, as it allows families to take part in studies in a more naturalistic home setting, at a time convenient to them, improving the success of data acquisition (Ross-Sheehy et al., 2021; Zaadnoordijk et al., 2021). It equally helps researchers to acquire larger sample sizes within a shorter time by testing participants in parallel (Semmelmann et al., 2017; Chouinard et al., 2019; Rhodes et al., 2020; Zaadnoordijk et al., 2021). A recent unmoderated online study with 8-12-year-old children confirmed that participant attrition, task comprehensibility, technological difficulties, and parental interference pose no major challenges in such experiments (Nussenbaum et al., 2020). Available research thus suggests that online methods are a feasible and helpful tool for studying developmental questions.

Despite the new advances in online developmental research, the feasibility of paradigms for testing infants online remains to be comprehensively evaluated (Rhodes et al., 2020; Zaadnoordijk et al., 2021). First findings from the comparison of in-lab and online paradigms (i.e., looking time, preferential looking, sequential decision making, and verbal reports) suggest that

developmental phenomena can be examined not only in a laboratory setting but also through online experiments with infants (Scott et al., 2017; Kominsky et al., 2021a; Smith-Flores et al., 2021) and children (Sheskin and Keil, 2018; Nussenbaum et al., 2020; Kominsky et al., 2021a; Lo et al., 2021). Specifically, preferential looking time paradigms are widely used in infancy research (Dunn and Bremner, 2017), thus hold promise for online implementation. In such paradigms, infants observe two stimuli presented side-by-side on a computer screen while their gaze is recorded with an eye-tracker and/or a video camera to measure the total amount of time spent looking at each stimulus during a given time interval (Chouinard et al., 2019). The stimulus with longer fixations is considered to be preferred or novel/surprising by the infant participant (Aslin, 2007, 2012; Semmelmann et al., 2017).

Prior research with infants and children showed that looking time paradigms can be successfully applied online using webcam-based video recording (Scott and Schulz, 2017; Semmelmann et al., 2017; Tran et al., 2017; Lo et al., 2021; Smith-Flores et al., 2021). With the latest surge in the development of online experiment platforms, it is becoming even easier for researchers to conduct infant-looking time studies remotely and in an unmoderated fashion (Rhodes et al., 2020). Families can simply use their own computer and webcam to record and upload eye-tracking and video data on online experiment platforms such as Lookit (Scott and Schulz, 2017) and Labvanced (Finger et al., 2017). However, reproducing in-lab measurement accuracy and data quality with a webcam can pose a considerable challenge with infant participants even for video recordings, not to mention eye-tracking (Chouinard et al., 2019; Zaadnoordijk et al., 2021).

Eye-tracking is a prevalent method in infancy research for studying the development of perceptual and cognitive processes, as it allows to objectively and non-invasively measure gaze locations of young infants. Yet, the quality of eye-tracking data obtained from infants is often lower compared to data from adults because of lower accuracy and precision, as well as increased data loss (Gredebäck et al., 2009; Wass et al., 2014; Hessels and Hooze, 2019). Even though eye-tracking still works reasonably well with infants in laboratory conditions, webcam-based eye-tracking involves limitations such as poor image quality and uncontrolled experimental conditions (i.e., infant positioning, lighting in the room, and presence of distractors; Wass, 2016; Zaadnoordijk et al., 2021). To our knowledge, there are no published studies that have used webcam-based eye-tracking with infants. However, methodological advances in online research with adults demonstrated that webcam-based eye-tracking systems can obtain data in comparable quality to data gathered in a traditional lab setting (Xu et al., 2015; Papoutsaki et al., 2016; Bott et al., 2017; Semmelmann and Weigelt, 2018), and even smartphones can reach the accuracy of mobile eye-trackers (Valliappan et al., 2020). Collecting eye-tracking data online entails higher variance, a lower sampling rate (Gagné and Franzen, 2021), and increased experimental time, but shows no significant differences in spatial accuracy compared to in-lab recordings for adult data (Semmelmann and Weigelt, 2018). Nonetheless, no research to date has

directly compared webcam-based and in-lab eye-tracking data from infants.

The aim of the current study was to examine whether webcam-based eye-tracking is a feasible method to assess infants' basic perception abilities, specifically the detection of audio-visual temporal synchrony. Temporal synchrony is the amodal information that enhances the perception of integrated stimuli from multisensory input and its detection emerges early in development (Lewkowicz, 1996). Although infants as young as 4 months can detect temporal asynchrony between simple audio-visual stimuli (e.g., a bouncing ball hitting the ground; Provasi et al., 2017), the ability to detect audio-visual asynchrony of complex stimuli, such as a person dancing to instrumental music, emerges only between 8 and 12 months (Hannon et al., 2017). However, these findings seem to contradict evidence suggesting that infants' musical abilities are present from birth (Winkler et al., 2009) and their sensitivity to synchrony in early social interactions emerges at 3-4 months (Murray and Trevarthen, 1986; Feldman, 2012). Based on the above, infants may be more likely to determine asynchrony between audio-visual stimuli when these stimuli are familiar and socially meaningful to them. The preferential-looking paradigm applied in this study was designed to investigate whether infants can detect audio-visual asynchrony between stimuli that are simple and familiar (i.e., infant being bounced to music) compared to stimuli that are complex and less familiar to them (i.e., person dancing to music).

Using this paradigm, the present study set out to evaluate the feasibility of online infant eye-tracking in direct comparison to in-lab eye-tracking, especially in the case of preferential looking. We assessed 4-6-month-old infants in a between- and within-subjects design. One group was tested online using webcam-based eye-tracking and video recording, whereas the other group was assessed in the laboratory with conventional eye-tracking and video recording. Online and in-lab eye-tracking data were compared in terms of data quality, infants' viewing behavior, and experimental effects.

First, we expected that eye-tracking data quality will be similar between the two groups, based on previous results from the adult literature revealing no significant difference in spatial accuracy between in-lab and online eye-tracking (Semmelmann and Weigelt, 2018). Since preferential looking time paradigms with infants require lower spatial accuracy in terms of gaze behavior, online eye-tracking could be a feasible tool to provide comparable data with in-lab eye-tracking. As measures of data quality, we assessed eye-tracking calibration quality, sampling frequency, missing data quantity, and average task and trial duration. Calibration quality is a crucial measure to compare the accuracy and precision of online and in-lab eye-tracking and can be evaluated quantitatively or qualitatively (Nyström et al., 2013; Dalrymple et al., 2018). Sampling frequency (the number of times the eyes' positions are registered per second) also needs to be carefully contrasted between the two methods. While lab-based eye-tracking devices have a typical sampling rate of 500-1000 Hz, online sampling rates may only reach 30 Hz due to technical limitations of the participant's device and the eye-tracking algorithm itself (Gagné and Franzen, 2021).

Missing data quantity or data loss (the relation between the expected number of gaze samples recorded by the eye-tracker and the actual number delivered) typically ranges from 2 to 20% in in-lab eye-trackers (Cuve et al., 2021). Data loss can be even higher in infant eye-tracking due to the many short periods of data loss, which cannot be attributed to infants looking away or blinking (Hessels and Hooze, 2019). Thus, the data acquired from online eye-tracking with infants need to be assessed for data loss. Finally, comparing the average duration of the eye-tracking task between methods can be informative as it may reveal more frequent pauses or a lower level of concentration in the online setting, further affecting data quality (Semmelmann and Weigelt, 2018).

We also contrasted the two methods on video data quality including completeness, frame rate per second (fps), brightness, resolution, and usability based on previous studies with adults (Semmelmann and Weigelt, 2018) and infants (Scott and Schulz, 2017; Scott et al., 2017; Semmelmann et al., 2017). The measures of video completeness and usability can be indicative of participants' compliance with instructions as well as the suitability of their experimental set-up for online recording. Sufficient frame rate per second and resolution are important for accurate video annotation (Scott and Schulz, 2017) that can complement eye-tracking data analysis (Fraser et al., 2021). Luminance or brightness of video recordings can also impact the ability of the eye-tracking algorithm to detect the participant's face during calibration and the experimental task (Semmelmann et al., 2017; Fraser et al., 2021) as well as the feasibility and pace of video data annotation. Additionally, parental interference was assessed from the videos and compared between groups to account for the potential influence of the familiar home environment.

Next, we hypothesized that viewing behavior of infants is independent of the method used, meaning that eye-tracking and video recording can capture infants' gaze behavior to rather large areas of interest (AOIs) uniformly in both experimental settings. As eye-tracking and video recording are applied complementarily in in-lab preferential looking studies to provide accurate data, the same should be achievable by online eye-tracking complemented with video recording. Finally, we anticipated that experimental effects would manifest in better asynchrony perception (higher looking time differences) in case of simple vs complex stimuli in accordance with the findings of Provasi et al. (2017), irrespective of the method used. To explore whether the online study reached a more diverse population, the in-lab and online samples were contrasted on caregiver education level. Caregivers' education level in both groups was further compared with parental education levels in the generic population.

To conclude, in the current unmoderated online study, we aimed to compare the feasibility of in-lab and online infant eye-tracking in a preferential-looking paradigm, which assessed infants' audio-visual synchrony perception. Our study provides a direct comparison of in-lab and webcam-based eye-tracking data from infants who completed an identical looking time paradigm in two different settings – in the laboratory or online at home.

MATERIALS AND METHODS

In line with open science practices, the in-lab study was pre-registered on AsPredicted¹. As full in-lab data assessment ($n = 30$) could not be completed due to the current COVID-19 pandemic, data collection was continued online, which necessitated a comparison of data quality between the in-lab and online procedures and thus motivated the current paper.

Participants

Overall, 91 infants in the age range of 4–6 months participated in the study, 45 in the laboratory and 44 online. Participants were recruited from a database of volunteers, our research unit's website (<https://kinderstudien.at/>), via online advertisements on social media (Facebook, Twitter), and an online participant recruitment platform (<https://kinderschaffenwissen.eva.mpg.de/>). Participation criteria included no prior knowledge of the Hungarian language to ensure that the audio stimuli in the experimental task were not previously known to participants. We have included 38 infants in the final sample: 18 from the in-lab procedure ($M = 4.9$ months; $SD = 16$ days; 8 girls) and 20 from the online procedure ($M = 5.2$ months; $SD = 31$ days; 6 girls). We excluded 27 in-lab participants due to fussiness ($n = 5$), incomplete video data ($n = 6$), or because of insufficient calibration ($n = 16$). From the online participants, we excluded 23 infants due to several attempts of the experimental task ($n = 1$), no calibration error data ($n = 14$), or because of high calibration error (more than 5 degrees of visual angle; $n = 9$). The unusually high attrition rates both in-lab and online were partly due to technical issues and partly because the study constituted the first infant eye-tracking study conducted in a newly established laboratory and the relative inexperience of the newly trained experimenters. Notably, it was also the first online eye-tracking study conducted by the authors, so level of (in-)experience was in fact similar for both data assessment modes. All included infants were typically developing and born at term, with a gestation period of at least 37 weeks. The 10-min APGAR score (a simple numerical assessment of a newborn's health performed 1, 5, and 10 min after birth; Apgar, 1966) was greater than 9/10 ($n = 30$), indicating little to no complications after birth. Mothers' age averaged 32.34 years ($SD = 4.27$) and 79% of them had a university degree. All infants came from middle- to upper-class families based on parental education. Infants had no auditory or visual impairments as assessed by maternal report. Written informed consent was obtained from all infants' legal guardian before participation in the laboratory or online. The study was approved by the Ethics Committee of the University of Vienna, Austria. Participation in the laboratory was remunerated.

Design and Stimuli

One group of infants was tested in the laboratory and the other group was assessed online, while both groups completed the same experiment. The experimental task consisted of two conditions (simple and complex) and a total of 12, 23-s-long

trials. Each trial was preceded by a 3-s animated attention getter (a spinning star) accompanied by an infant-friendly sound to direct infants' attention to the center of the screen. In each trial, infants were presented with visual stimuli, namely, two side-by-side videos, one of which was synchronous, while the other one was asynchronous with an auditory stimulus. The areas of the two videos shown on the screen constituted the two AOIs for later gaze data analysis. AOI size was 609×1080 frame units in both settings; in the in-lab setting, this was equivalent of 12×29.2 cm, whereas in the online setting, the actual size depended on the screen size of the participant's device. The complexity of both the visual and auditory stimuli was manipulated according to the condition. In the simple condition, the audio-visual stimuli were two videos of an unfamiliar infant being bounced rhythmically up and down to a Hungarian children's song sung by a female voice with infant-directed singing (**Figure 1A**). In the complex condition, the stimuli were two videos of an unfamiliar woman dancing (based on Hannon et al., 2017) to the same Hungarian children's song sung by a duet of female voices with instrumental orchestra accompaniment (**Figure 1B**). In both conditions, synchrony between the auditory and visual stimuli was altered by manipulating the meter. As the original auditory stimulus had a meter of 4/4, in the synchronous videos the movements were performed in 4/4 meter (with stress on the first beat), while in the asynchronous videos the movements were performed in 3/4 meter (with stress on the first beat). The presentation order of the conditions and the position of the two videos (synchronous and asynchronous) on the screen (left/right) were pseudorandomized across participants using four different trial sequences (lists) to avoid order and position biases. Each list consisted of a total of 12 trials administered in 3 blocks. Each block consisted of four trials, two simple and two complex ones (**Figure 2**). The trials within a list were alternated based on condition (simple/complex), to avoid consecutive repeats of trials from the same condition. Two lists started with a simple trial, while the two other lists started with a complex trial. The position of the synchronous stimulus (left/right) was pseudorandomized across trials and lists. Within each list, for six trials the synchronous stimulus was shown on the left, while for the other six trials, on the right. The total duration of the experimental task was approximately 6 min (excluding the time for initial calibration; and in the online procedure, the time for saving the participant's video data after each trial). For the in-lab study, the experiment was programmed in the software Experiment Builder (Version 2.1.1, SR Research Ltd.), whereas for the online study, it was implemented with the online experiment platform LabVanced (Finger et al., 2017).

In-Lab Data Acquisition

Participants sat on an experimenter's ($n = 13$) or a caregiver's ($n = 5$) lap approximately 60 cm distant from the presentation monitor (17 inches, 37.6×29.2 cm, resolution: 1850×1090 pixels). To avoid distraction during the eye-tracking task, infants and experimenters/caregivers were seated behind a wall separating them from the other experimenter(s) and/or caregiver and the rest of the laboratory room. Infants' binocular gaze data were recorded using an EyeLink 1000 Plus (SR Research Ltd.) eye-tracking system, arm mount with remote mode. The

¹<https://aspredicted.org/ck4za.pdf>

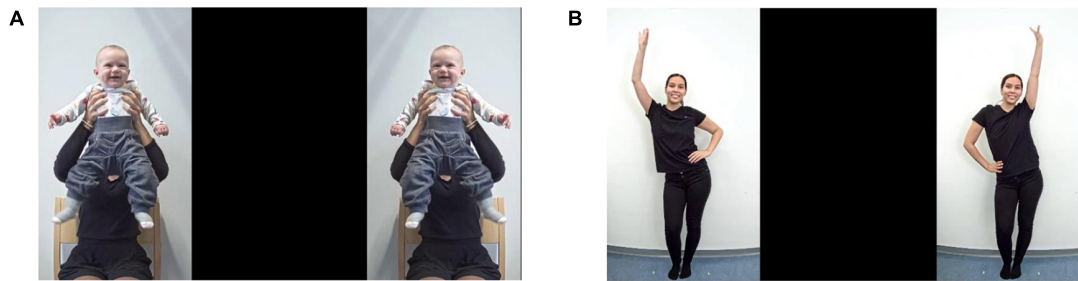


FIGURE 1 | Simple and complex stimuli during an experimental trial. **(A)** In the trials of the simple condition, the audio-visual stimuli were two side-by-side videos of an unfamiliar infant being bounced rhythmically to a simple version of a children's song. **(B)** In the trials of the complex condition, the stimuli were two side-by-side videos of an unfamiliar woman dancing to the complex version of the same children's song. In each trial of both conditions, one video was synchronous while the other one was asynchronous with the song; and the positions of the two videos (left/right) were pseudo-randomized across trials.

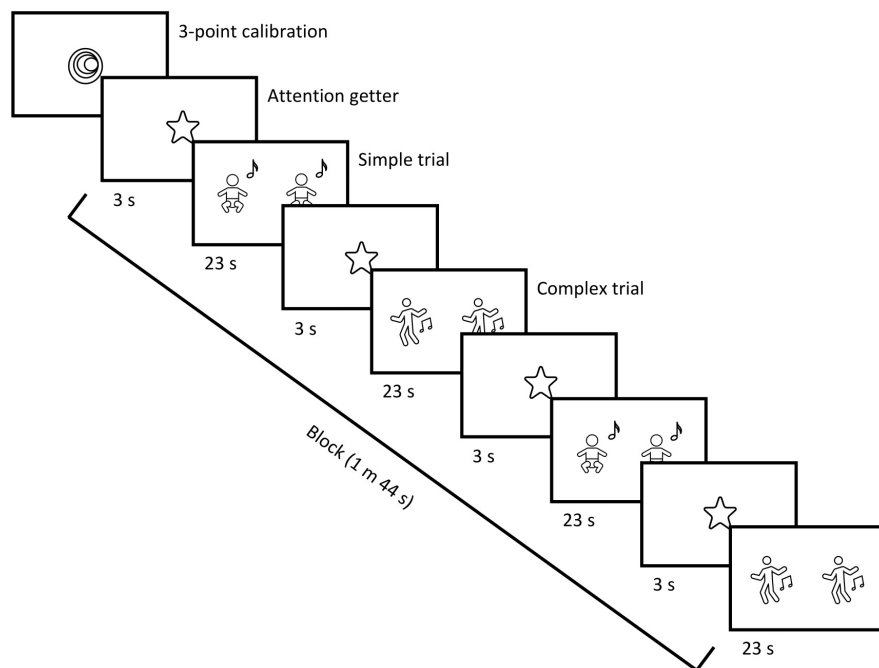


FIGURE 2 | Experimental block (duration: 1 m 44 s). Each block consisted of four trials, two simple and two complex ones. Participants saw 12 trials in total, administered in three blocks (corresponding to one list). Trial duration was 23 s, and each trial was preceded by a 3-s-long, infant-friendly attention getter (spinning star). Trials within a block and list were alternated based on condition (simple/complex), to avoid consecutive repeats of trials from the same condition. The position of the synchronous stimulus (left/right) was pseudorandomized across trials and lists. Prior to the first block, a three-point calibration (with a spinning spiral) was performed.

eye-tracking camera had a 16 mm/1:14 infant lens, with a 940 nm illuminator. The presentation computer had an Audio Stream Input/Output (ASIO; Steinberg Media Technologies GmbH) compatible sound card, which assured high synchrony of audio-visual stimuli presentation. The sound was delivered *via* external stereo speakers placed behind the presentation monitor. The eye-tracking system was controlled using the software EyeLink 1000 Plus (Version 1.0.12, SR Research Ltd.) on a second computer out of infants' sight. The light in the room was dimmed and turned on just behind the participant during the task. Lighting conditions across participants were kept constant by closing the window blinds in the room. Caregivers were instructed to be silent and

not to interfere with the experiment both if they were holding the infant on their lap or if they were observing the experiment from behind the separator wall. The person (experimenter or caregiver) on whose lap the infant was sitting, was instructed not to move and avoid speaking to, or in other ways interfering with the infant during the experiment. First, the focus of the eye-tracker camera was manually adjusted while infants saw an infant-friendly animation (a crab) moving on the screen. Next, a three-point bilinear calibration was performed as recommended for younger infants (Farroni et al., 2007; Di Giorgio et al., 2012; Bardi et al., 2015). Calibration stimulus consisted of an infant-friendly animation (a spinning spiral) accompanied with a

twinkling sound to draw the infant's attention toward the screen. For validation, the animation was shown again in the center of the screen with a circle-shape AOI around it (size: 198×192 frame units, diameter: 5 cm). The experimenter visually inspected if the infant's gaze was within this area for the calibration to be accepted. If the calibration was not successful, another attempt was performed. During the experiment, the infant's face was video recorded at 60 fps with a Sony Action Camera HDR-AS200V (Sony Corporation) positioned under the presentation screen, and with an associated live view remote. Upon completion of the task, caregivers were requested to fill out a self-report questionnaire to provide basic demographic information about the infant (age, gender, APGAR scores, language(s) used in the family, musicality, caregivers' age and education level, presence of any auditory or visual impairments).

Online Data Acquisition

The online version of the experimental task was hosted on the LabVanced experiment platform, a JavaScript web application that offers a graphical user interface to implement behavioral research studies online *via* an internet browser while providing users with full experimental control (Finger et al., 2017). The link to the study together with an access password was sent to the participants' caregivers in an individualized invitation email, upon providing written informed consent for participation *via* our research unit's website. Participation was possible with several devices, including computers with the operating systems Linux, Mac OS, and Windows, as well as Android Tablets and iPads. The option to use smartphones was not enabled, as their small screen size would not be comparable with the in-lab presentation screen. Minimum screen resolution was set to 600×600 pixels. Supported browsers included Chrome, MS Edge, and Opera. The study was available in English and German, according to the participant's choice. Prior to starting the task, caregivers saw on-screen instructions asking to make sure their internet download speed is minimum 10–16 MB/s, to complete the experiment in a quiet room with no bright light sources behind them, and to wear sunglasses to prevent the webcam detecting their own eyes instead of the infant's. They were also instructed not to move and avoid speaking to, or in other ways interfering with the infant during the experiment. Additionally, caregivers were advised to continue with the task if the infant was comfortable even when not attending to all trials of the experiment. Further, instructions were provided about pre-programmed button-press commands with regard to ignoring head-pose checks (a built-in eye-tracking feature of the platform), taking a break during the experiment, skipping the task to move forward to the caregiver questionnaire, or stopping the study entirely at any time. A sound-check was also implemented: caregivers were asked to play a short, infant-friendly audio sample before starting the task, to make sure the volume is comfortably set for the infant.

Regarding positioning, caregivers had to set up the device on a table, sit on a chair in front of the screen and hold their infant on their lap leaning against their upper body, approximately 60 cm from the screen. To help with correct positioning, participants' webcam was activated prior to the task to display the infant's

seating position on the screen, while asking caregivers to make sure the infant's face can be clearly seen in the center of the screen. Before the calibration procedure was deployed, caregivers were asked (a) to check if the infant's head position is recognized by a virtual mask (a built-in eye-tracking feature of the platform), (b) to ensure that the infant is looking at the screen, and (c) to avoid moving the screen or the webcam from this point onward. Next, a nine-point, infant-friendly calibration was performed for 60 s. Calibration stimuli consisted of infant-friendly graphics of animals shrinking in size until fully disappearing into one calibration point after another of a nine-point grid shown on the screen. Each stimulus was accompanied by an appropriate animal sound, in order to draw the attention of the infant toward the screen. Upon completion, calibration data were saved while an infant-friendly video (a cat spinning on a record player) was shown for approximately 30 s. If the calibration was not successful, another attempt was performed. The experimental task was identical to the one in the in-lab procedure. During each trial, the infant's face was video recorded *via* the participant's own webcam by the in-built video recording feature of LabVanced at approximately 25–30 fps (based on hardware specifications of the individual webcams) and with a fixed upload speed of 512 kbit/s. No audio recordings were made to allow for better stimuli presentation and recorded video data quality. Following the experimental task, caregivers were requested to fill out the same self-report demographic questionnaire as in the in-lab procedure, implemented in an online format on LabVanced. Additional to the in-lab survey, caregivers were asked to provide information regarding the device type (computer, laptop, and tablet) used for the experiment, the type of their operating system, and their screen size and resolution. Caregiver reports about experienced technical issues were also collected here (i.e., missing sound, lagging videos, unstable internet connection, long waiting times, and other issues), plus they were asked if they skipped the head-pose check during the study. Eye-tracking, video, and demographic data were initially recorded on the LabVanced platform and were exported by experimenters after participants completed the task.

Data Preprocessing and Analysis

Demographic Data

Demographic data were collected from caregivers after the experimental task in the laboratory as well as online by a self-report questionnaire. Caregivers were asked to provide data on the infant's age, gender, APGAR scores, language(s) used in the family, musicality, caregivers' age and education level, and presence of any auditory or visual disorders. Musicality was assessed *via* a questionnaire, which included five-point Likert scale questions ($n = 4$) about the frequency of the infant listening to music, singing, making music together with the caregiver (i.e., 1 = very frequently; 5 = never), as well as caregiver musicality (i.e., 1 = very musical; 5 = not musical at all). It also contained dichotomous questions ($n = 9$) about musical routines (singing during bedtime routine, play situations, comforting, other situations), infants' musical education, and parental music practice (playing on an instrument, singing in a choir, and for

both: doing it professionally or as a hobby) (i.e., 1 = yes; 2 = no). For overall musicality, a composite score was calculated based on the sum of these answer scores (lower scores indicating higher musicality).

To rule out any potential effect of the demographic background variables on between-group differences, we compared infants' age, gender, musicality, and multilingualism, as well as caregivers' age and education level between the two groups. In addition, caregivers' education levels in both groups were further compared with caregiver education levels in the generic population of Austrian families (Austrian Federal Ministry of Health, 2016). For this last analysis, participants from another country than Austria were excluded ($n = 4$).

Video Coding

All videos were micro-coded (frequency, duration) for parental interference and infants' viewing behavior using Datavyu, a free, open-source video coding software (Version 1.37²; Lingeman et al., 2014). Interference was coded when an infant was visibly distracted by a caregiver who interfered by talking to, stroking, or moving the infant; moving her own arms and/or legs or the infant's arms and/or legs to the beat of the music; or pointing to the screen. Following the video annotation procedure applied in a prior study with infants conducted on Labvanced (Benavides-Varela and Reoyo-Serrano, 2021), infants' viewing behavior was coded as time spent looking to the AOI on the screen (left and right stimuli videos), to the middle of the screen, and away from the screen. One experimenter coded all data. To establish inter-rater reliability, 22% and 15% of randomly chosen in-lab and online videos (respectively) were independently coded by a trained research assistant for viewing behavior and interference. As no interference events could be identified, reliability was only assessed for viewing behavior. Cohen's kappa (Cohen, 1960, 1968) was calculated between the coding of the two raters and resulted in $\kappa = 0.94$ for the in-lab and $\kappa = 0.89$ for the online sample, indicating sufficiently high inter-rater agreement.

Data Quality

First, to gain a more detailed overview on the experimental settings in the online sample, participants' device type, operating system and browser type, screen size and resolution, as well as the number of times they attempted to start the study was explored. The number of excluded infants was also compared between groups. The frequency of technical issues with the experimental setup or other issues reported by the experimenter (in-lab) and by the caregiver (online) as well as the number of attempted trials were compared between groups.

Second, eye-tracking data quality was assessed for both groups, specifically calibration quality, sampling frequency, and missing data quantity. Raw gaze data recorded with the in-lab eye-tracker were extracted using the software EyeLink Data Viewer (Version 3.1.1, SR Research Ltd.), whereas raw gaze data from Labvanced were readily downloadable in a comma-separated values file for each participant. To assess the level of calibration quality, in-lab eye-tracking session data were

assessed for the level of calibration. Since no validation procedure with average error recording could be performed, a categorical evaluation was made. Calibration quality was considered high if both eyes were calibrated, fixations fell in the AOI of the attention getter shown during the validation-like event, and no recalibration was required during the task. Quality level was assessed as medium if all these criteria were met, but only one eye could be calibrated; or in case both eyes were calibrated but recalibration was needed. Low calibration quality was concluded if only one eye could be calibrated, and recalibration was required. As a measure of online calibration quality, the Labvanced eye-tracking algorithm recorded an average calibration error value for each participant in frame units (e.g., a 100-unit error is equivalent to 2.5 degrees of visual angle/cm). Calibration quality was evaluated high in case the error was under 2.5 degrees of visual angle, medium if it was between 2.5–3.75, and low if it was between 3.75 and 5 (Dalrymple et al., 2018). Sampling frequency (the number of gaze positions returned by the eye-tracker per second), and the percentage of missing samples were compared between groups. Average task duration was also calculated from the start of the first trial until the end of the last trial based on UNIX timestamps recorded by the in-lab eye-tracker and the online platform and compared between groups. The same analysis was performed for average trial duration, which was calculated as the differences of the trial-level start and end timestamps averaged over trials.

Finally, video data quality was contrasted between groups. Videos from both in-lab and online participants were assessed for video usability. Videos were usable if they were available and complete for all trials the infant had completed. Video data quality between groups was compared on fps and resolution using the software FFmpeg (Version 4.4, Tomar, 2006), as well as on brightness, which was extracted for a randomly selected snapshot from each video in MATLAB (Version R2018b).

Viewing Behavior

To investigate the accuracy of each method, we assessed if infants' viewing behavior recorded by the eye-tracker matched with respective gaze durations coded from the videos. That is, we compared infants' trial-level fixation durations (in-lab) or gaze durations (online) to the two AOIs recorded by the eye-tracker with respective looking times to both AOIs coded from the videos within and between groups. Next, the number of valid trials with sufficient eye-tracking data quantity (defined as data recorded for at least 70% of the video duration) was determined and contrasted between groups. For these valid trials, trial-level fixation durations (in-lab) or gaze durations (online) to the synchronous AOI (relative to the total looking time to both AOIs in the trial) were compared between participants' eye-tracking and video recordings within group.

For calculating the in-lab fixation durations, nearby fixations that were shorter than 200 ms were merged. For each participant, fixation durations to AOIs (right, left) were extracted separately for left and right eye samples (where available). Final data were obtained through a custom MATLAB script that calculated fixations durations, independently from the eye sampled. Fixation durations were calculated considering both eyes, so that

²<https://datavyu.org/>

when the fixation start and end time of the two eye samples were not overlapping (i.e., a fixation was detected only from one eye), the duration of this fixation was calculated from the available eye sample data. This approach allowed to obtain fixation data even for time intervals when one eye was not detected by the eye-tracker (i.e., due to the infant turning the head while still looking at the screen). Overlapping samples recorded from the two eyes at the same time point were expected to fall in the same AOI due to the large size of our AOIs and due to the fact that the movements of infants' two eyes are conjugated. For extracting the gaze durations to AOIs recorded online, time differences between consecutive eye samples were calculated. Each sample recording contained the x and y gaze position coordinates that allowed the assignment of the respective AOI (right, left, middle, away) to the sample *post hoc*. Samples with missing gaze position coordinates and/or timestamps were discarded.

To analyze the experimental effect, we calculated infants' trial-level relative looking times to the synchronous and asynchronous stimuli by dividing the time spent looking at a certain AOI with the total gaze duration to both AOIs during a trial. Relative looking times were calculated based on the fixation/gaze durations (in-lab/online) recorded by the eye-tracker/webcam, as well as the looking times coded from the videos. These looking time variables were tested against chance in each condition within each group and then contrasted between groups and conditions separately to test for infants' audio-visual synchrony perception while accounting for any potential effect of the method used.

Statistical Analyses

All statistical analyses were carried out in the free, open-source statistical software JASP (Version 0.14, JASP Team, 2021) and RStudio (Version 1.3.1093, RStudio Team, 2020). For certain data visualizations, the Raincloud-shiny online plotting application was also used (Allen et al., 2021). To account for any between-group differences moderated by demographic background variables, we performed between-group comparisons for infants' age and musicality with Welch's *t*-tests; for infants' gender and multilingualism with chi-square tests; for caregivers' age with two-sample *t*-tests; and for caregivers' education level with Mann-Whitney *U* tests. Caregiver education level proportions in our sample were compared with respective education levels in the generic population of Austrian families using *z*-tests.

For the analyses of data quality, first we compared the frequency of technical issues with the experimental setup and the number of excluded infants between groups applying chi-square tests. Then the number of attempted trials was compared between groups using a two-sample *t*-test. Regarding eye-tracking data quality, the frequency of high-, medium-, and low-level calibration quality was descriptively compared between the in-lab and online sample (due to no average validation error recordings were available in the in-lab sample). Total and trial-level sample count, as well as average task and trial duration were compared between the two groups by two-sample *t*-tests. A Mann-Whitney *U* test was used to compare the percentage of missing samples between groups. Video data quality between groups was descriptively compared on fps and resolution and contrasted on brightness using a Welch's *t*-test.

Infants' trial-level fixation/gaze durations (in-lab/online) to the two AOIs were compared with the respective looking times to both AOIs coded from the videos within group using a paired-sample *t*-test and between groups with a two-sample Welch's *t*-test. Based on the first analysis, the number of valid trials with sufficient eye-tracking data quantity (defined as data recorded for at least 70% of the video duration) was determined and contrasted between groups with a Welch's *t*-test. At this point, infants with less than two valid trials per condition were excluded from further analyses of the eye-tracking data (in-lab: $n = 6$; online: $n = 13$). For infants with a sufficient number of valid trials, these trials were extracted. For these valid trials, trial-level fixation/gaze durations (in-lab/online) to the synchronous AOI (relative to the total looking time in the trial) were compared between participants' eye-tracking and video recordings within group using Wilcoxon rank sum tests.

To test if infants' relative looking times were different from chance level (50%) in each condition within group, one-sample *t*-tests were performed. To estimate the effects of group and condition on the relative looking time spent (per trial) on the synchronous stimulus (proportion values), a Generalized Linear Mixed Model was used (GLMM; Baayen, 2008) with a Beta distribution and logit link function. Group and condition were included into the model as fixed effects, individual infant as random effect, and condition within individual infant as random slope. The variable condition was manually dummy coded and centered before being included into the slope applying an R function kindly provided by Roger Mundry. The model was fitted in R using the package GLMMTMB (Brooks et al., 2017) for relative looking times from eye-tracking. Then the same model was fitted on relative looking times from the video recording.

For the eye-tracking data, the model encompassed 150 proportion values, taken from 19 infants (in-lab: $n = 12$, online: $n = 7$) out of two groups (in-lab/online) during two conditions (simple/complex; with min. two valid trials per condition). In order to check for collinearity among the predictors, we also determined Variance Inflation Factors (VIF; Field, 2005) based on a standard linear model, lacking the interaction and the random effects. This revealed collinearity to be no issue (maximum VIF: 1). With a dispersion parameter of 0.89, the response was not overdispersed. For the video looking time data, the model encompassed 454 proportion values, taken from all 38 infants in the sample out of two groups (in-lab/online) during two distinct conditions (simple/complex; all trials). Collinearity and overdispersion were not present (VIF: 1; dispersion parameter: 1.03). We expected an effect of condition but not group on the relative looking time to the synchronous stimulus both for eye-tracking and video recording.

RESULTS

Demographics

To rule out any potential effect of the demographic background variables on between-group differences, we first compared infants' age, gender, musicality (lower scores indicated higher musicality thus were reverse-scored for data visualization),

and multilingualism, as well as caregivers' age and education level between the two groups. There were no statistically significant differences between groups in terms of infants' age, $t(29.24) = -1.29$, $p = 0.21$ (**Figure 3A**); gender, $\chi^2(1, n = 38) = 0.85$, $p = 0.36$ (**Figure 3B**); multilingualism, $\chi^2(1, n = 38) = 0.07$, $p = 0.79$ (**Figure 3C**); musicality, $t(26.27) = 1.84$, $p = 0.08$ (**Figure 3D**); maternal age, $t(36) = -0.31$, $p = 0.76$ (**Figure 3E**); paternal age, $t(36) = -0.43$, $p = 0.67$ (**Figure 3F**); maternal education level, $W = 139$, $p = 0.1$ (**Figure 3G**); and paternal education level, $W = 153.5$, $p = 0.39$ (**Figure 3H**).

Next, maternal, and paternal education levels in both the in-lab and online samples were compared with the proportions of caregivers with respective education levels in the generic population of Austrian families. For this analysis, participants with another country of origin than Austria were excluded ($n = 4$ in the online sample). In the remaining sub-sample, maternal education level was significantly higher in the online than in the in-lab group, $W = 103$, $p = 0.05$; whereas paternal education level did not differ between groups, $W = 121$, $p = 0.38$. Thus for the following analyses, data from the two groups were assessed separately for maternal education level but were collapsed on paternal education level across groups.

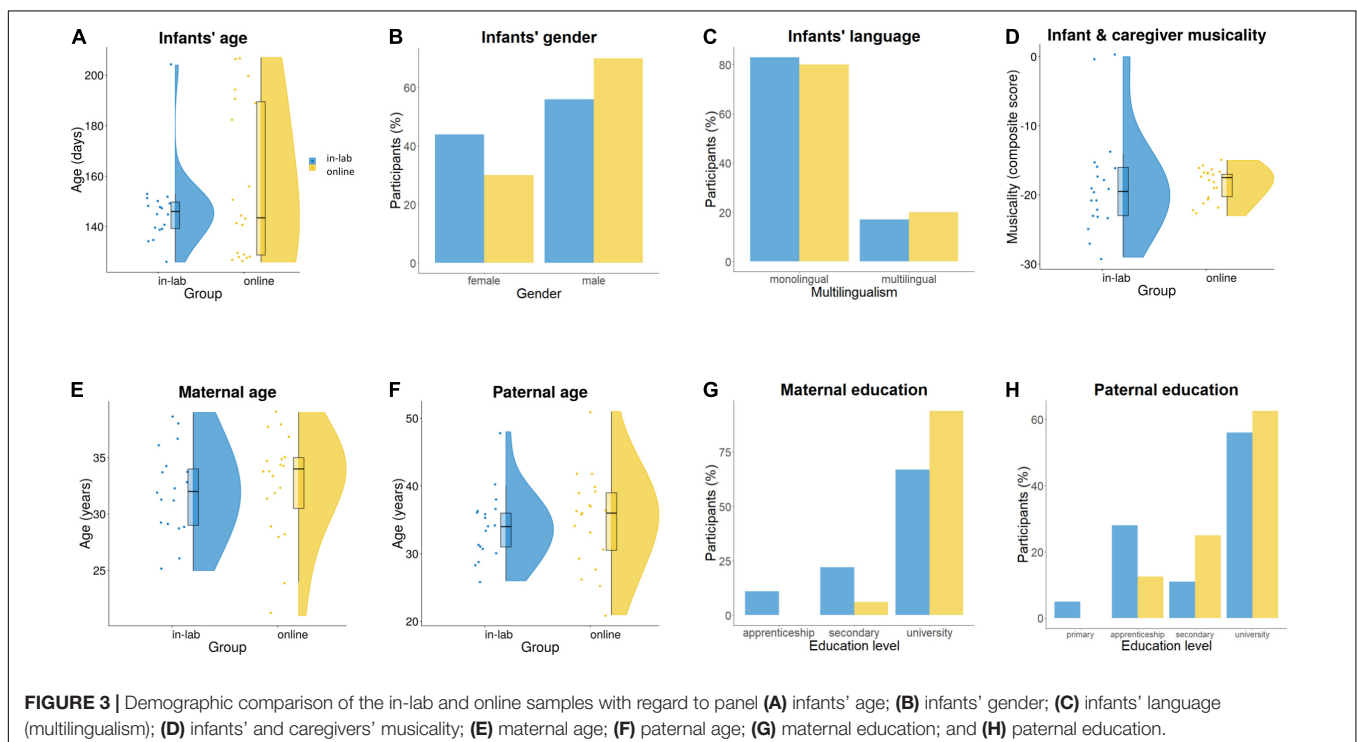
The proportion of mothers with university, college, or university-related education was significantly higher in our in-lab and online samples (67%; 94%) than in the generic population (16%), $z = 5.75$, $p < 0.001$; $z = 8.25$, $p < 0.001$. The proportion of mothers with apprenticeship was not significantly different in the in-lab group (11%) compared to the generic population (30%), $z = -1.75$, $p = 0.08$; but was significantly lower in the online group (0%) compared to the generic population, $z = -2.61$,

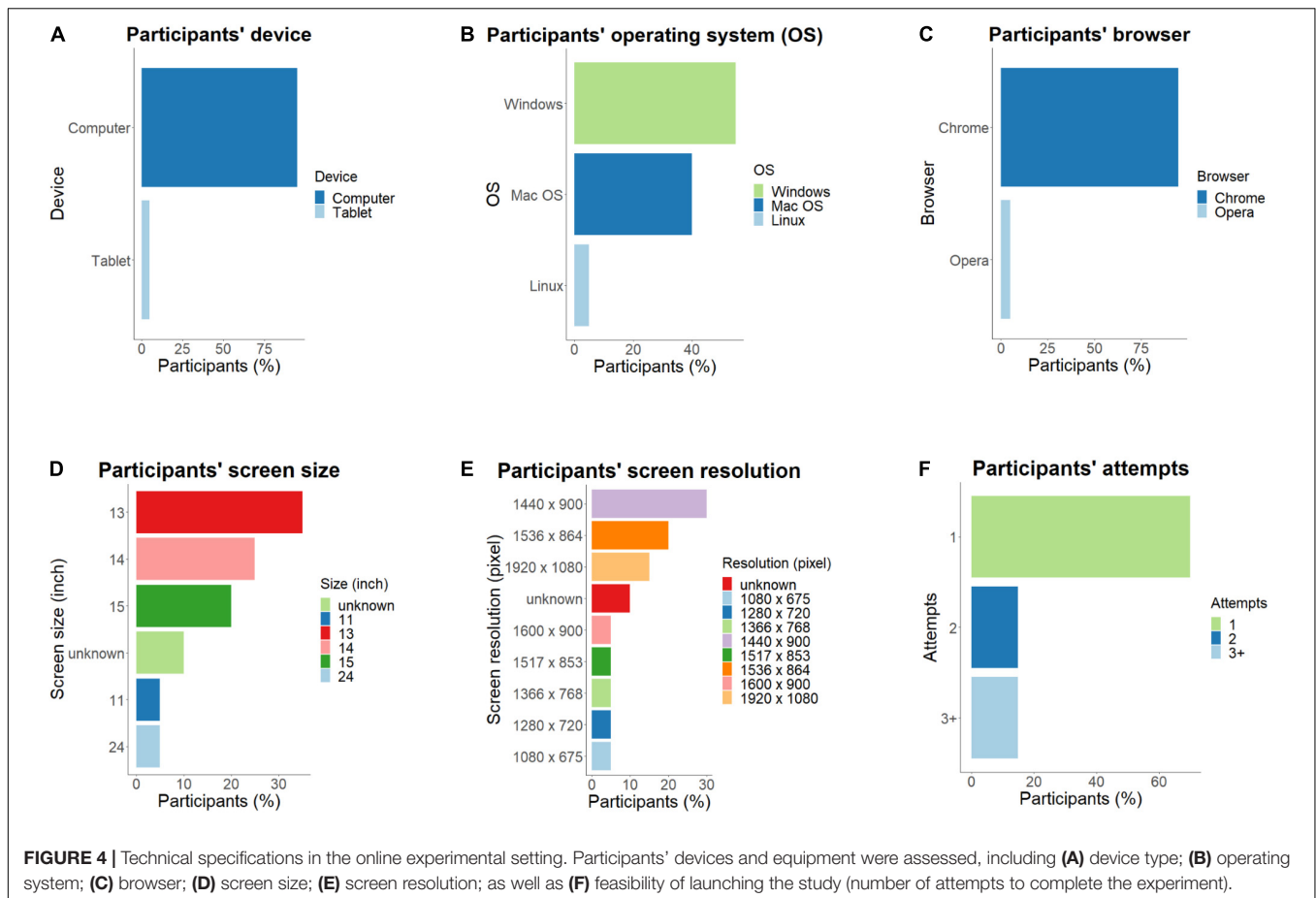
$p < 0.01$. Maternal secondary level education was equally frequent in our in-lab and online samples (22%; 6%) and in the generic population (18%), $z = 0.44$, $p = 0.66$; $z = -1.25$, $p = 0.21$. The proportion of mothers with primary level education was significantly lower in our in-lab and online samples (0%; 0%) than in the generic population (19%), $z = -2.05$, $p = 0.04$; $z = -1.93$, $p = 0.05$. The proportion of fathers with university, college, or university-related education was significantly higher in our overall sample (59%) than in the generic population (18%), $z = 6$, $p < 0.001$. The proportion of fathers with apprenticeship was significantly lower in our sample (20%) compared to the generic population (42%), $z = -2.57$, $p = 0.01$. Paternal secondary level education was equally frequent in our sample (12%) and in the generic population (14%), $z = -0.33$, $p = 0.74$. The proportion of fathers with primary level education was not significantly different: 5% in our sample and 11% in the generic population, $z = 0.73$, $p = 0.47$.

Data Quality

Experimental Settings

To gain an overview on the technical aspects of the experimental setting in the online sample, participants' device type, operating system and browser type, screen size and resolution, as well as the frequency and nature of technical issues were explored. The majority (95%; $n = 19$) of online participants used a computer to complete the experimental task, while only 5% ($n = 1$) used a tablet (**Figure 4A**). With regard to the operating system (OS), 55% of the participants had Windows ($n = 11$), 40% Mac OS ($n = 8$), and only 5% Linux ($n = 1$) (**Figure 4B**). The majority of participants (95%; $n = 19$) ran

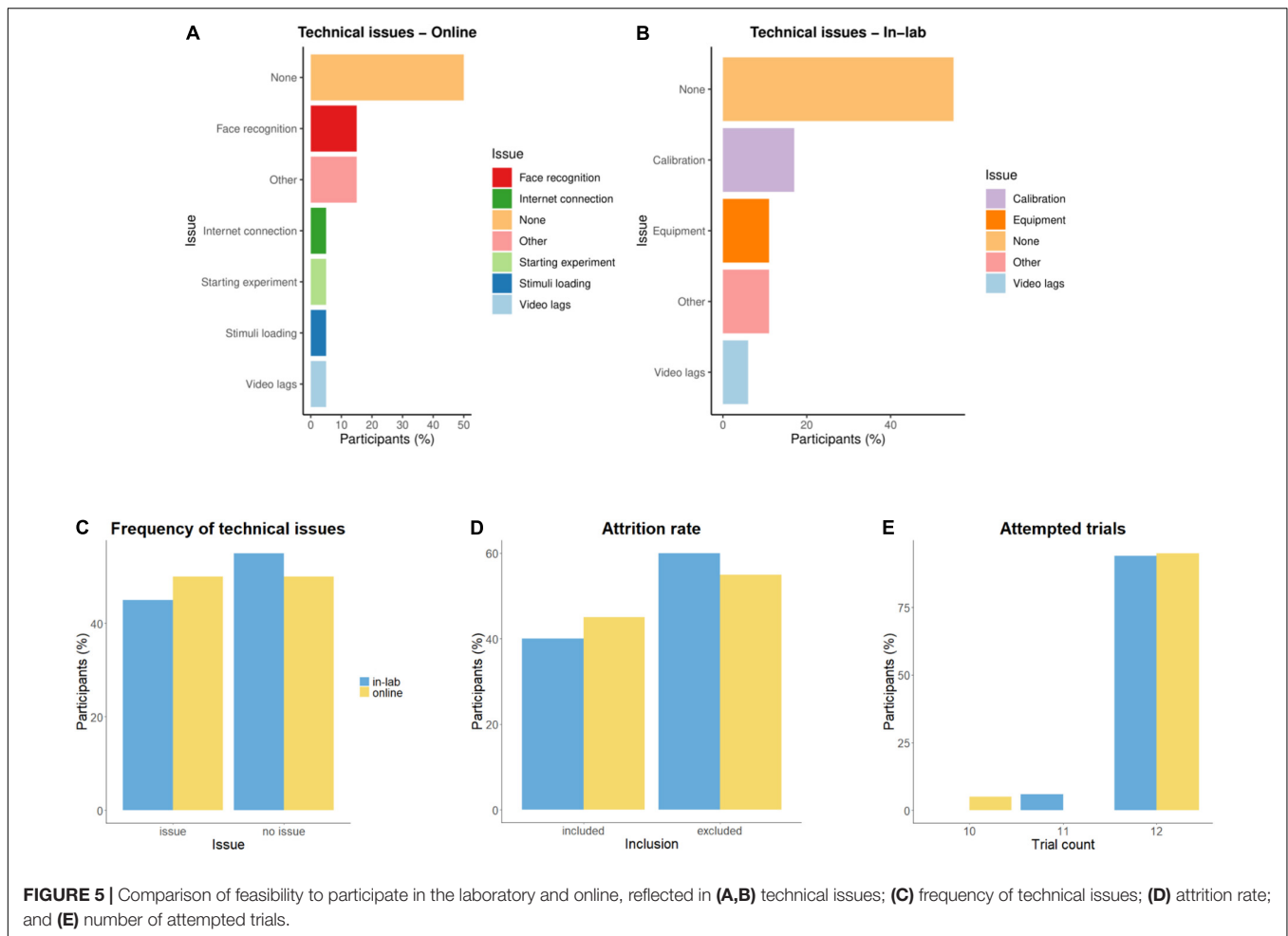




the experiment from a Chrome browser, and 5% from Opera ($n = 1$) (Figure 4C). Participants' screen size varied between 11 and 24 inches, whereas resolution ranged between 1080×675 and 1920×1080 pixels (Figures 4D,E). For comparison, in-lab participants were presented with the experimental task ran on a Windows computer, on a 17-inch screen with a resolution of 1850×1090 pixels. Since there was no variance in the in-lab group regarding these variables, we can conclude that device type and operating system were mostly identical, while screen size and resolution were more varied in the online group. The in-lab procedure did not rely on an internet connection; thus no browser was used. Regarding the number of attempts online participants made to start the study, 70% ($n = 14$) managed to complete the study at the first attempt, while 15% ($n = 3$) at the second, and 15% ($n = 3$) at the third attempt (Figure 4F). For participants who attempted the study more than once, the experimental task had not been always initiated, thus they likely encountered issues already at the phase of the instructions and/or the eye-tracking calibration. On these initial, unsuccessful attempts, no eye-tracking and video data were recorded.

With regard to the frequency and nature of technical or other issues, 50% ($n = 10$) of the online participants reported some sort of problem: 15% ($n = 3$) had difficulties with infants' face recognition; for 5% ($n = 1$), the experiment only started on the second attempt; 5% ($n = 1$) experienced internet

connection problems; 5% ($n = 1$) faced long waiting times due to stimuli loading; 5% ($n = 1$) had occasional video lags; and 15% ($n = 3$) indicated other issues (i.e., the infant became fussy/inattentive after some time) (Figure 5A). Interestingly, data available from participants regarding skipping the head-pose check ($n = 15$) – a built-in eye-tracking feature added to LabVanced shortly after the study started – show that only 20% ($n = 4$) used this option, but among these participants, three reported no technical issues, while one reported internet connection problems. Based on this, we assume that face recognition issues as reported by 15% of participants were not critical enough to make caregivers deactivate the head-pose check entirely (for the whole duration of the task), thus could be disregarded when assessing data quality. However, caution should be exercised when analyzing eye-tracking data for those infants whose head-pose check was skipped during the study. In the lab, technical or other issues were reported for 45% ($n = 8$) of the participants: for 17% ($n = 3$), several calibration attempts were necessary to achieve sufficient calibration; for 11% ($n = 2$), computer issues occurred (i.e., low sound, hardware/software errors); 6% ($n = 1$) had occasional video lags; and 11% ($n = 2$) had other issues (fussiness/inattention) (Figure 5B). There were no statistically significant differences between the two groups in the frequency of technical or other issues during the experiment, $\chi^2(1, n = 38) = 0.12, p = 0.73$ (Figure 5C), in the



number of excluded infants, $\chi^2(1, n = 38) = 0.38, p = 0.54$ (Figure 5D), or in the number of attempted trials, $t(36) = 0.38, p = 0.71$ (Figure 5E).

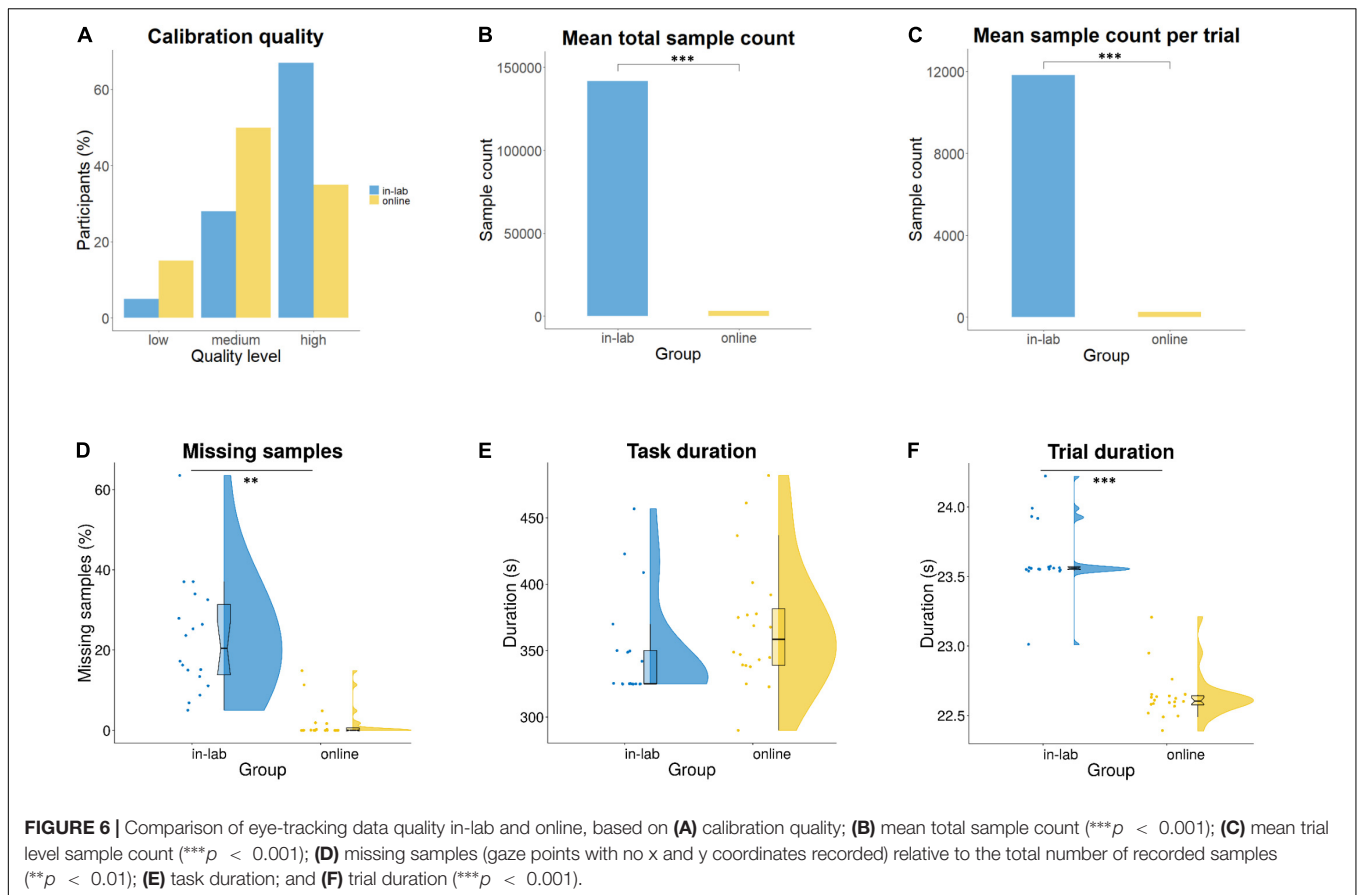
Eye-Tracking Data Quality

Regarding calibration quality, in the in-lab sample, 67% of infants had high-, 28% medium-, and 5% low-level quality, whereas in the online sample, 35% of infants had high-, 50% medium-, and 15% low-level quality (Figure 6A). Sampling frequency (the number of gaze positions returned by the eye-tracker per second) was set to 500 Hz in the in-lab procedure and defined as 20–28 Hz for the online eye-tracking algorithm (a gaze point recorded in every 30–50 ms) on the experiment platform. For the online data, the actual sampling rate was calculated by dividing the total number of samples collected during all the trials with the overall task duration. The actual sampling rate for the online group was 11.52 Hz on average ($SD = 6.1$). There was a significant difference in the total sample count, as well as in the trial level sample count between groups, $t(36) = 263.55, p < 0.001$; $t(36) = 264.63, p < 0.001$. Total and trial level sample counts were higher in the in-lab than in the online group (Figures 6B,C). The percentage of missing samples (gaze points with no x and y coordinates recorded) relative to the total number of recorded

samples was significantly higher in the in-lab, than in the online group, $W = 351, p < 0.01$ (Figure 6D). In the in-lab setting, 23.13% ($SD = 13.98$) of samples were lost on average, whereas in the online sample, this occurred only for 1.76% ($SD = 3.98$) of the samples. However, sampling frequency in the laboratory was 500 Hz, while online it was only 11.52 Hz on average. No significant difference was found in the average task duration between groups, $t(36) = -1.31, p = 0.19$ (Figure 6E). However, there was a significant difference in the average trial duration, $t(36) = 14.05, p < 0.001$ (Figure 6F). Average trial duration was measured as 23.63 s ($SD = 0.25$) in the lab and 22.64 s ($SD = 0.17$) online.

Video Data Quality

The video coding procedure confirmed that videos were recorded for all in-lab and online participants. All videos were complete and usable: they included recording of all attempted trials and allowed for infant gaze coding. In-lab videos uniformly had a resolution of 1920×1080 pixels and 59.94 fps as were recorded with the same camera. Online videos had lower resolution: 1280×720 ($n = 19$) or 640×480 ($n = 1$) pixels and a lower average frame rate of 23.56 fps ($SD = 7.78$) (Figures 7A,B). Brightness values extracted for randomly



selected video snapshot images were not significantly different between groups, $t(19.6) = -0.94$, $p = 0.36$ (Figure 7C).

Viewing Behavior

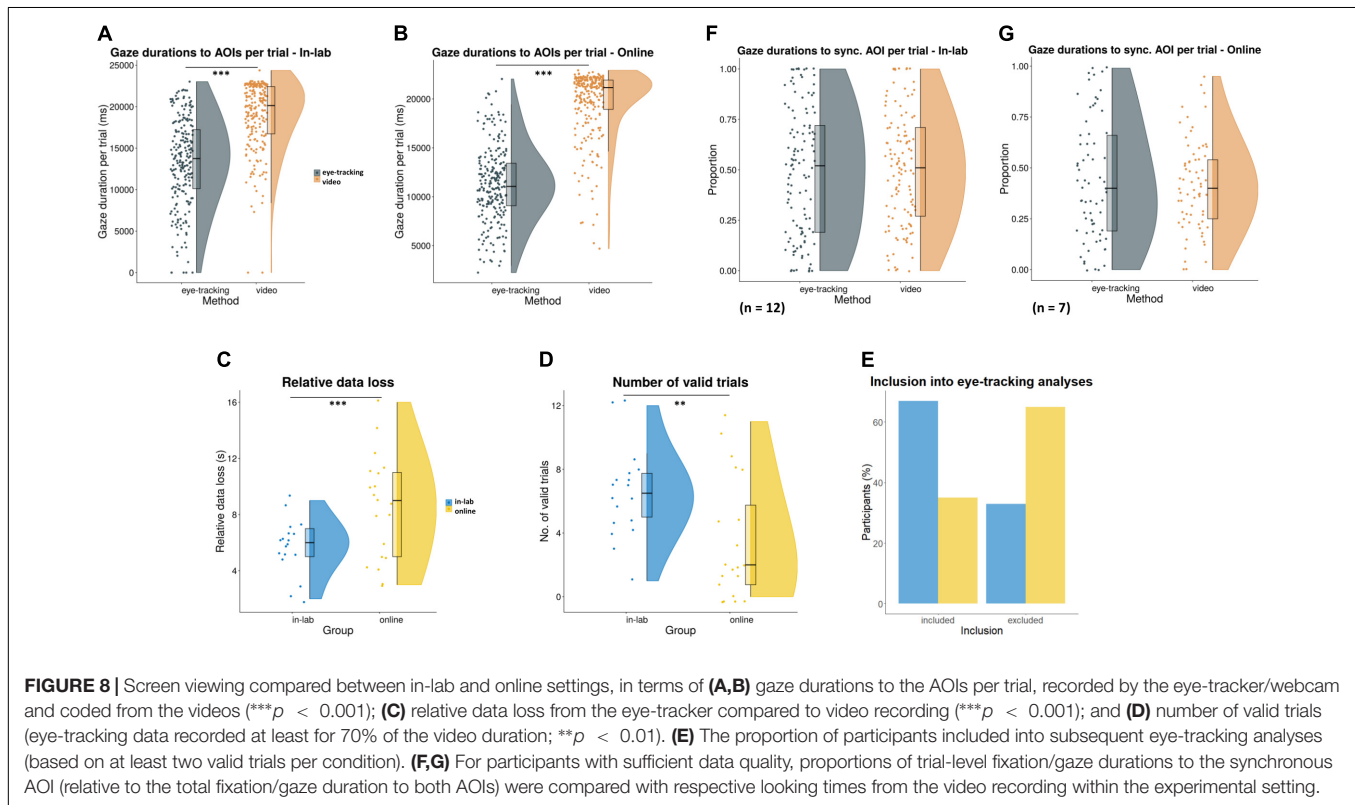
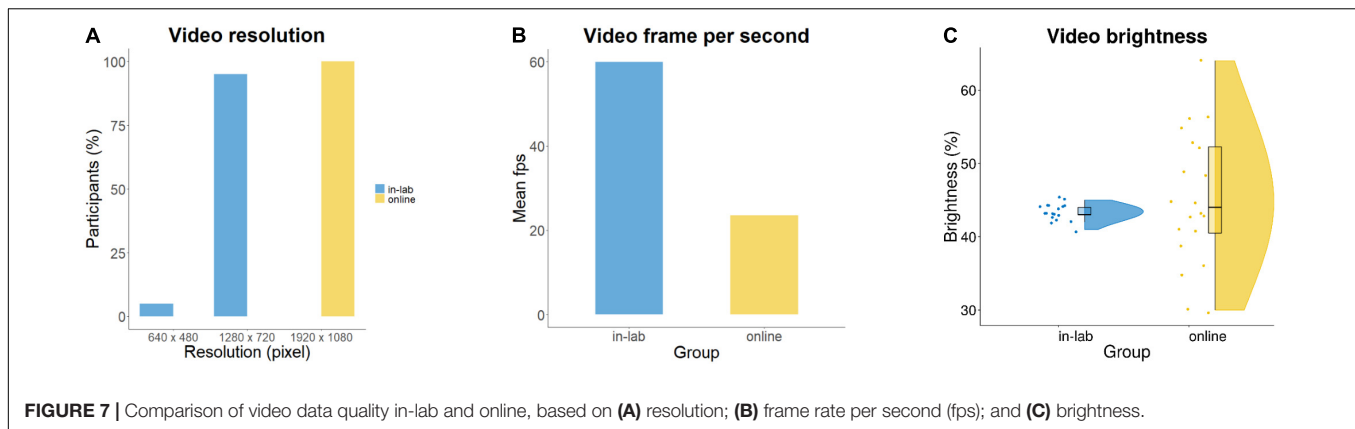
Screen Viewing

Within the in-lab group, infants' trial-level fixation durations to the two AOIs recorded by the eye-tracker were significantly lower than the respective looking times to both AOIs coded from the videos, $t(215) = -19.25$, $p < 0.001$ (Figure 8A). The same results were found for the online group: infants' trial-level gaze durations to the two AOIs captured by the eye-tracker were significantly lower than the respective looking times coded from the videos, $t(237) = -27.17$, $p < 0.001$ (Figure 8B). This relative data loss from the eye-tracker compared to video recording was also significantly higher in the online group than in the in-lab group, $t(452) = -6.39$, $p < 0.001$ (Figure 8C). Based on the within-group analysis, the number of valid trials with sufficient eye-tracking data quantity (data recorded for at least 70% of the video duration) was on average 6.5 out of 12 in the in-lab and 3.5 out of 12 in the online sample. Infants in the in-lab group had a significantly higher number of valid trials compared to infants in the online group, $t(34.93) = 2.83$, $p < 0.01$ (Figure 8D). Overall, 67% of in-lab ($n = 12$) and 35% of online participants ($n = 7$) had enough trials to be included in the subsequent eye-tracking data analyses (Figure 8E).

For in-lab participants with a sufficient number of valid trials, the proportions of the trial-level fixation durations to the synchronous AOI (relative to the total fixation duration to both AOIs in the trial) were not significantly different from respective looking times from the video recordings, $W = 2247$, $p = 0.43$ (Figure 8F). In case of online participants, results (with gaze durations) were identical, $W = 1147.5$, $p = 0.63$ (Figure 8G).

Preferential Looking Effects

The analysis of eye-tracking data showed that relative looking time spent at the synchronous stimulus was significantly different from chance level in the online group in the complex condition, $t(27) = -2.07$, $p < 0.05$. This comparison was not significant for the simple condition in the online sample, $t(27) = -0.59$, $p = 0.56$, nor for any of the conditions in the in-lab sample, $t(43) = -1.97$, $p = 0.34$ (simple); $t(49) = -0.23$, $p = 0.82$ (complex). Relative looking time as coded from videos to the synchronous stimulus differed significantly from chance only in the simple condition in the online group, $t(118) = -2.26$, $p = 0.03$, but not in the complex condition in the online group, $t(118) = -1.54$, $p = 0.13$, nor in any of the conditions in the in-lab group, $t(107) = -0.92$, $p = 0.36$ (simple); $t(107) = -0.8$, $p = 0.43$ (complex) (Figure 9). Group and condition and their interaction as fixed effects had no significant impact on the trial-level relative looking time to the synchronous stimulus neither for the eye-tracking data (Table 1) nor for video recordings (Table 2).

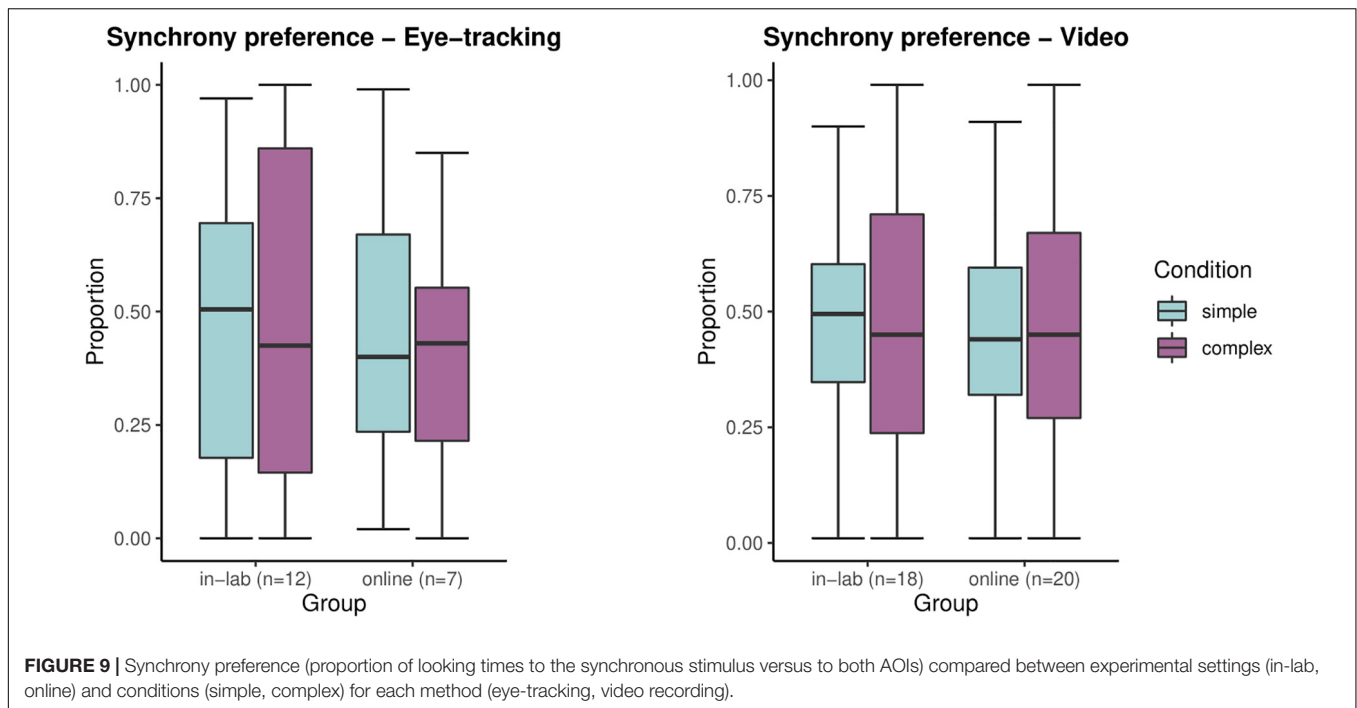


DISCUSSION

In summary, this study provides first insights into the feasibility of online infant eye-tracking, especially in the case of preferential-looking paradigms. A direct comparison of webcam-based and in-lab eye-tracking and video data is essential to assess whether online data collection methods with infants can generate reliable and reproducible results. Further, our aim was to offer methodological and practical considerations to researchers designing and conducting online eye-tracking experiments with infants, an avenue becoming ever more important to developmental research in recent times. First, we discuss the advantages and challenges of both methods with regard to data acquisition and data quality. Then we outline our results

on infants' viewing behavior and the assessed experimental effect. Finally, we evaluate the potential of online studies for reaching diverse participant groups, based on the demographic characteristics of our sample.

Data acquisition was performed both in-lab and online as part of a first eye-tracking study in a newly established laboratory. Attrition rates were thus higher (60% in the laboratory and 52% online) than usually reported with infants at this age (e.g., 48% in Frank et al., 2009; 33% in Michel et al., 2021), likely due to experimenter inexperience as well as technical issues (calibration errors in both groups) and infants' fussiness. Attrition rates were not significantly different between groups. Online attrition could be further explained by the fact that the study was conducted in an unmoderated format, thus



caregivers could not access immediate assistance for technical issues from experimenters. This limitation was compensated by the advantage that participants could complete the study online at any time convenient to them. From the excluded online participants, 32% of caregivers could not complete the study until the end, thus calibration error recording was missing for these infants. A limitation was that this recording was performed for the initial participants at the end of the experimental task, which did not allow the error value to get recorded for participants who could not complete the study. Our recommendation is thus to perform this recording of calibration error right at the start of the experimental task. From the included online participants, 30% of caregivers attempted to complete the task more than one time. These issues indicate that some of the online participant families faced challenges with maintaining infants' attention or could have lacked the required hardware and internet connection speed for the study. The most common issues reported by 50% of the online participants included difficulties with infants' face recognition (a built-in feature for online eye-tracking), starting the experiment, internet connection problems, and long waiting times during stimuli loading and recorded video file upload. In the laboratory, technical issues affected 44% of included participants and consisted of insufficient calibration, equipment issues, stimuli video lags, and infant fussiness or inattention. Our findings indicate that the frequency of technical issues and the number of attempted trials were not significantly different between the in-lab and online samples. We also found no events of experimenter or caregiver interference during the completion of the experimental task in the laboratory or online, suggesting that our caregiver instructions for avoiding interference with the infant were efficient in both cases. Therefore we conclude that experimental conditions for recording eye-tracking and

video data from infants online are comparable to the ones in the laboratory, in line with findings from previous studies with infants, older children, and adults (Scott and Schulz, 2017; Scott et al., 2017; Semmelmann et al., 2017; Tran et al., 2017; Lo et al., 2021; Smith-Flores et al., 2021). When setting up online experiments with infants, we encourage for sufficient study planning, preparation of detailed caregiver instructions, and frequent exchange with the technical support of online experiment platforms to ensure the experimental conditions are kept as identical as possible with those in the laboratory. Based on experiences from the present study, we agree with Zaadnoordijk et al. (2021) that unmoderated data collection online allows families to participate in studies from the comfort of their home at a convenient time, ensuring a similar success of data acquisition for researchers as in the laboratory. By testing participants in parallel, we were able to acquire a sufficient sample size online, which would not have been as easily achievable in the laboratory due to the current worldwide pandemic situation. Additional technical assistance for online participants depending on experimenters' availability and capacities could further increase study completion success rate and thus final sample size.

We assessed data quality for in-lab and online recordings for both eye-tracking and video data. With regard to the in-lab eye-tracking calibration quality, a limitation to point out is that no validation procedure with average calibration error recording could be performed. Therefore, we only conducted a categorical comparison of calibration quality between the in-lab and online groups and found that 67% of in-lab participants had high calibration quality, while this was only the case for 35% of online participants. Medium calibration level was achieved for 28% in-lab and 50% online participants. Further studies with a similar focus should aim to record average calibration

TABLE 1 | Results of the GLMM of trial-level relative looking times to the synchronous stimulus measured with eye-tracking, with estimates, standard errors, z-values, and confidence intervals (CIs).

| Relative looking times (eye-tracking) | | | | | |
|---------------------------------------|----------|------------|---------|-----------------|------------------|
| | Estimate | Std. Error | z-Value | Lower CI (2.5%) | Upper CI (97.5%) |
| (Intercept) | −0.01 | 0.17 | −0.03 | −0.33 | 0.32 |
| Group (online) | −0.36 | 0.28 | −1.29 | −0.9 | 0.19 |
| Condition (simple) | −0.15 | 0.24 | −0.62 | −0.62 | 0.32 |
| Group * condition | 0.46 | 0.4 | 1.17 | −0.31 | 1.24 |

TABLE 2 | Results of the GLMM of trial-level relative looking times to the synchronous stimulus measured with video recording, with estimates, standard errors, z-values, and confidence intervals (CIs).

| Relative looking times (video recording) | | | | | |
|--|----------|------------|---------|-----------------|------------------|
| | Estimate | Std. Error | z-Value | Lower CI (2.5%) | Upper CI (97.5%) |
| (Intercept) | −0.04 | 0.1 | −0.41 | −0.23 | 0.15 |
| Group (online) | −0.12 | 0.13 | −0.91 | −0.38 | 0.14 |
| Condition (simple) | −0.03 | 0.14 | −0.24 | −0.3 | 0.23 |
| Group * condition | 0.05 | 0.19 | 0.25 | −0.32 | 0.41 |

error for more exact comparisons between in-lab and online eye-tracking accuracy. As infant-friendly calibration on the online experiment platform was preconfigured, it could have contributed to the lower calibration quality levels in the online group. Developing more customizable calibration procedures for online infant eye-tracking studies could allow researchers to prepare a personalized procedure more suited for the age group they assess. While the sampling rate of the eye-tracker was 500 Hz in the laboratory, the actual online eye-tracking sampling rate was altogether 12 Hz, lower than expected from the online experiment platform (20–28 Hz), a finding which is in line with results from Semmelmann and Weigelt (2018). As no previous infant eye-tracking studies, to our knowledge, have been reported thus far, we speculate that this lower sampling rate could be due to participants' hardware specifications, technical issues, or infants' excessive movement due to fussiness. Online experiment platforms will need to increase sampling rate in future to ensure higher precision of webcam-based eye-tracking, especially when assessing infant participants. However, even in the case of higher sampling rates, the limitations inherent to participants' own hardware specifications would remain unchanged. Such difference in sampling frequency between in-lab and online eye-tracking poses a considerable limitation for comparing data with high precision from the two methods. Total and trial level sample count were both higher in the in-lab than in the online group due to the higher sampling frequency of the in-lab eye-tracker. Interestingly, the percentage of missing samples relative to the total number of samples was significantly higher in the in-lab than in the online group: in the laboratory, 23% of all samples were lost on average, whereas in the online sample, only 2%. This finding likely indicates technical issues with the in-lab eye-tracking data recording (i.e., calibration problems, infants'

fussiness), but could also suggest a higher level of attention retention in the online participant group due to completing the task at home. Average experimental task duration was uniform between the in-lab and online groups, whereas average trial duration was slightly longer in the laboratory, likely due to the marginally different allocation of timestamps to trial start and end times by the two eye-tracking systems.

We also contrasted the methods of eye-tracking and video coding on video usability (Scott and Schulz, 2017), overall experimental duration and video data quality including completeness, frame rate per second (fps), brightness, and resolution (Semmelmann et al., 2017). In both samples, all our recorded videos were complete and usable. The video recordings included all attempted trials and allowed for infant gaze coding. In the in-lab, but not in the online sample, attrition due to missing video data occasionally still occurred. While online video data acquisition is automatically deployed by the experiment platform, the necessity to control video recording manually in the laboratory leaves a higher chance for experimenter error. This could be avoided by using built-in, automated video recording combined with eye-tracking also in laboratory procedures. In-lab videos had a higher fps and resolution than online videos, allowing only a less accurate comparison of video data between the two methods. Caregiver instructions in the online sample ensured that lightning conditions were kept under sufficient control, resulting in no significant differences (but higher variability) in brightness between the videos of the in-lab and online samples.

Next, we investigated infants' viewing behavior in terms of screen viewing and experimental effects. Results from the analysis of in-lab screen viewing showed that infants' trial-level fixation durations to the AOIs recorded by the eye-tracker were significantly lower than respective looking times to the same AOIs coded from the videos. This finding is in line with results from a previous study contrasting data loss from eye-tracking compared to video coding data of children's viewing behavior (Venker et al., 2020), and is likely explained by the high relative number of missing samples in the in-lab group, which raises concerns about the accuracy of the eye-tracking measurement. Identical results could be seen in the online sample: infants' trial-level gaze durations to the AOIs captured by the eye-tracker were significantly lower than the respective looking times coded from the videos. Despite a low relative number of missing samples in the online group, the video recording had an average fps of 24, whereas eye-tracking sample frequency was only 12 Hz. This difference likely explains the mismatch between eye-tracking and video data. Additionally, we cannot entirely rule out that the video coding conducted by two independent raters still lacked sufficient accuracy, contributing to these results. Future studies could overcome this limitation by establishing a more extensive pilot study prior to data collection to ensure higher eye-tracking and video data accuracy for each method. Moreover, we found a significantly higher relative data loss from the eye-tracker as opposed to the video recording in the online- compared to the in-lab sample. This result can be explained by the lower sampling rate and calibration accuracy of the online eye-tracker and the lack of its precise fixation duration recording (i.e., only

gaze coordinates and timestamps but no fixation durations are recorded). This is additionally supported by the fact that video data quality was similar between methods due to a similar fps between the laboratory camera and participants' webcams, as well as the potentially higher level of infant attention in a familiar home setting. We further assessed the number of valid trials with sufficient eye-tracking data quantity (data recorded for at least 70% of the video duration). Our results show that infants in the in-lab group had a significantly higher number of valid trials compared to infants in the online group. This was likely due to the lower eye-tracking data quality in the online sample. As a future consideration, it could be worthwhile to include a higher number of trials into online infant eye-tracking experiments with an opportunity for caregivers to skip individual trials, not only the whole experimental task (as in our study). Overall, 67% of in-lab and 35% of online participants had enough valid trials (i.e., at least 2 per condition) to be included in the subsequent eye-tracking data analyses of screen viewing and experimental effects. For these infants, we first re-assessed accuracy between eye-tracking and video recording within group. For both in-lab and online participants, we found that the proportions of the trial-level relative fixation/gaze durations to the synchronous AOI were not significantly different from respective looking times coded from the videos, indicating better accuracy between eye-tracking and video recording in case of valid trials vs all trials in general.

We did not find a statistically significant and consistent effect of stimulus complexity or experimental setting (group) in our data. Our findings revealed that infants in the online group were able to distinguish between synchronous and asynchronous displays in the simple condition (when measured with video recording) and in the complex condition (when measured with eye-tracking). As these results are not in line with the main hypothesis and not consistent across methods, they call for further investigation. Based on results from our model, infants between 4 and 6 months of age in this sample did not detect asynchrony easier for simple stimuli than for complex stimuli. These findings are partially in line with work from Hannon et al. (2017), who found that infants between 5 and 8 months do not yet differentiate auditory mismatch between socially complex audio-visual stimuli. A novelty preference for such stimuli seems to develop between 8 and 12 months. For future studies in this direction, we suggest reducing stimuli complexity further and including older infants in order to extensively explore the development of the expected effect between 4 and 12 months of age. A larger sample size of 4–6-month-olds would also allow more accurate comparisons within this age range to account for potential developmental differences. Additionally, the examination of the role of active musical engagement and caregivers' musicality level may reveal individual differences in infants' perception of temporal synchrony.

With respect to the demographic composition of our sample, we can conclude that the in-lab and online participant groups were largely homogenous. There were no differences between the two sub-samples with regard to infants' age, gender, language, musicality, as well as caregivers' age and education level. When we compared caregivers' education level for Austrian participants

with the generic population of Austrian families, the proportion of caregivers with university-level education was significantly higher both in the in-lab and online groups than in the generic population. While caregivers' secondary level education in our sample was identical with the one in the generic population, levels of apprenticeship and primary education were significantly lower in some of the in-lab and online caregiver sub-samples (mothers/fathers) compared to the generic population. This finding contradicts recent claims in the literature that online research can reach larger sample sizes and increase participant diversity (Oakes, 2017; Lourenco and Tasimi, 2020; Zaadnoordijk et al., 2021). In the present study, participant recruitment for the online sample often relied on contacting families in our research unit's database and *via* personal connections. Extending participant recruitment by harnessing social media opportunities or by setting up collaborations with early childhood educators and versatile family networks may better ensure a more diverse sample. A further consideration is that online studies can only be run if families have the relevant hardware and stable internet connection. This entails limitations not only in terms of the sample characteristics but also the global application of these studies. Initiatives such as the ManyBabies Consortium (Zaadnoordijk et al., 2021) sets a promising example to tackle these issues by supporting cross-recruitment of participants across studies (in accordance with local ethics regulations) while facilitating an exchange of best practices among researchers.

Taken together, our study has several limitations. First, our sample sizes were rather small due to high attrition rates and the subsamples were homogenous in terms of caregivers' age and education. In terms of online participant recruitment, families with limited access to suitable hardware and steady internet connection had less opportunities to take part. Similarly, caregivers with a concern for their infant's exposure to excessive screen time may have also opted out from the online study. In the laboratory, the main limitations included experimenter inexperience and technical issues during data collection. Even though the online experiment platform was user friendly, setting up the study and acquiring the data were substantially affected by the novelty of the online experimental method. We needed to adapt our online data acquisition to the continuous development process of the online experiment platform (i.e., features for more accurate timestamp recording, skipping head position check, and recording the confidence for gaze points were only developed and added during the data acquisition process). Additionally, the available measures were not fully identical in the in-lab and online samples (e.g., exact calibration error values were not recorded in the laboratory to ensure a shorter, infant-friendly validation procedure; only durations between gaze points but not fixation durations could be recorded by the online eye-tracking algorithm), a limitation that prevented a more accurate comparison between the two methods. The accuracy of the calibration quality measure in the online sample could have been further hindered by the fact that 20% of online participants skipped a recalibration procedure (head-pose check) during the task to prevent infant fussiness. Studies with larger sample sizes could control for such participants and consider excluding them from further analyses. Moreover, stimulus presentation

timing in the online setting was not controlled for, but only assessed based on participant report (e.g., lags experienced in videos reported in the questionnaire). As there is a considerable variability in temporal precision between operating systems and browsers (Gagné and Franzen, 2021; Mathôt and March, 2021), future studies can circumvent this issue by recording the participant's screen and audio data. Yet, such recordings may add to the already high computational load on the participant's device. Limitations regarding eye-tracking data quality included a mismatch between the fixation (or gaze) durations recorded by the eye-tracker or webcam and the looking times coded from video recordings. As Venker et al. (2020) emphasize, this can lead to different patterns of results between eye-tracking and manual gaze coding, suggesting that the method used to analyze a particular research question could alter findings and the scientific conclusions that follow. Higher eye-tracking data accuracy could be ensured by further experimenter training in the laboratory, extensive technical support (both in-lab and online) and by providing participants with access to more suited devices for the online task (i.e., tablets). Video data coding could be further improved by using automated gaze coding (e.g., Fraser et al., 2021) and analyses software for video recordings. Regarding the analyses of experimental effects, a larger sample size with participants with a high number of valid trials may still alter the results presented here and such an analysis is among the goals of the authors to pursue in a subsequent study.

To conclude, our results indicate that online eye-tracking with infants is a promising avenue in developmental research and merits further exploration. However, the establishment of best practices for online data acquisition, data quality, and accuracy control, as well as analyses of data from a larger sample is essential (for a generic review, see Gagné and Franzen, 2021). Additionally, future studies aiming to assess the accuracy of online eye-tracking with adult and developmental populations could benefit from applying more challenging paradigms that require higher precision eye-tracking than preferential looking.

Our findings contribute to the first steps toward the development of online eye-tracking paradigms that could be applied widely with infant and child samples. Online eye-tracking and behavioral studies with infants can help to reduce data collection time and costs for researchers and participants (Tran et al., 2017; Semmelmann and Weigelt, 2018) and enhance replicability, reproducibility, and generalizability in developmental science (Rhodes et al., 2020; Visser et al., 2021). Our work will also inform future initiatives that aim to replicate in-lab studies with infants online and establish collaborations for large-scale, global online experiments (Frank et al., 2017; Byers-Heinlein et al., 2020; Sheskin et al., 2020; The ManyBabies Consortium, 2020; Zaadnoordijk et al., 2021).

REFERENCES

Allen, M., Poggiali, D., Whitaker, K., Marshall, T. R., van Langen, J., and Kievit, R. A. (2021). Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res.* 4:63. doi: 10.12688/wellcomeopenres.15191.2

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation. Raw video data are not readily available because of participant privacy.

ETHICS STATEMENT

The study involving human participants was reviewed and approved by the Ethics Committee of the University of Vienna. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

AB, GM, ME, and SH conceptualized the study. AB, GM, LF, and ME collected the data. ME preprocessed the in-lab eye-tracking data. LF preprocessed all video data. AB preprocessed the online eye-tracking data and analyzed all data with guidance from GM and SH. AB, GM, and ME wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Funding for this research, including open access publication fees, was provided by the University of Vienna.

ACKNOWLEDGMENTS

We wish to thank all children and families who participated in the study, without whom this research would not have been possible. We are grateful to the Department of Obstetrics and Gynecology of the Vienna General Hospital for support with our participant recruitment. We thank the Vienna Children Studies' research assistants and laboratory coordinator Liesbeth Forsthuber for their assistance with participant recruitment, data collection, and data preprocessing. We express our special thanks to Anna Matyk for video annotation and to Emese Égető, Luca Komáromi-Soós, Elisa Roberti, Myriam Spiegel, and Dóra Szalai for help with stimuli preparation. We also wish to thank the LabVanced and SR Research teams for continuous technical assistance, as well as Solveig Jurkat, Hanna Schleihau, Daniela Schmidt, and Takuya Yanagida for helpful suggestions regarding data collection and analyses.

Apgar, V. (1966). The newborn (APGAR) scoring system: reflections and advice. *Pediat. Clin. North Am.* 13, 645–650. doi: 10.1016/S0031-3955(16)31874-0

Aslin, R. N. (2007). What's in a look? *Dev. Sci.* 10, 48–53. doi: 10.1111/j.1467-7687.2007.00563.x

- Aslin, R. N. (2012). Infant eyes: a window on cognitive development. *Infancy* 17, 126–140. doi: 10.1111/j.1532-7078.2011.00097.x
- Austrian Federal Ministry of Health (2016). *Austrian Children's and Youth Health Report*. Vienna: Austrian Federal Ministry of Health.
- Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge, MA: Cambridge University Press.
- Bardi, L., Di Giorgio, E., Lunghi, M., Troje, N. F., and Simion, F. (2015). Walking direction triggers visuo-spatial orienting in 6-month-old infants and adults: An eye tracking study. *Cognition* 141, 112–120. doi: 10.1016/j.cognition.2015.04.014
- Benavides-Varela, S., and Reoyo-Serrano, N. (2021). Small-range numerical representations of linguistic sounds in 9- to 10-month-old infants. *Cognition* 213:104637. doi: 10.1016/j.cognition.2021.104637
- Bott, N. T., Lange, A., Rentz, D., Buffalo, E., Clopton, P., and Zola, S. (2017). Web camera based eye tracking to assess visual memory on a visual paired comparison task. *Front. Neurosci.* 11:370. doi: 10.3389/fnins.2017.00370
- Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., et al. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R. J.* 9, 378–400.
- Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., et al. (2020). Building a collaborative psychological science: lessons learned from ManyBabies 1. *Can. Psychol. Psychol. Can.* 61, 349–363. doi: 10.1037/cap0000216
- Chouinard, B., Scott, K., and Cusack, R. (2019). Using automatic face analysis to score infant behaviour from video collected online. *Infant. Behav. Dev.* 54, 1–12. doi: 10.1016/j.infbeh.2018.11.004
- Chuey, A., Lockhart, K., Sheskin, M., and Keil, F. (2020). Children and adults selectively generalize mechanistic knowledge. *Cognition* 199:104231. doi: 10.1016/j.cognition.2020.104231
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Measur.* 20, 37–46. doi: 10.1177/001316446002000104
- Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* 70, 213–220. doi: 10.1037/h0026256
- Cuve, H. C., Stojanov, J., Roberts-Gaal, X., Catmur, C., and Bird, G. (2021). Validation of Gazepoint low-cost eye-tracking and psychophysiology bundle. *Behav. Res. Methods* 2021:17. doi: 10.3758/s13428-021-01654-x
- Daghighi, S., Amini, M., Dodangeh, N., Hashemzadeh, M., Kiani Dehkordi, M., and Nekouei Shoja, N. (2020). 'Tele-observation' (with mobile phone) of infants discussed in online infant observation seminars during the 'new normal' of the Covid-19 pandemic. *Infant Observ.* 23, 7–15. doi: 10.1080/13698036.2020.1814842
- Dalrymple, K., Manner, M. D., Harmelink, K. A., Teska, E. P., and Elison, J. T. (2018). An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Front. Psychol.* 9:803. doi: 10.3389/fpsyg.2018.00803
- Di Giorgio, E., Turati, C., Altoè, G., and Simion, F. (2012). Face detection in complex visual displays: An eye-tracking study with 3- and 6-month-old infants and adults. *J. Exp. Child Psychol.* 113, 66–77. doi: 10.1016/j.cognition.2015.04.014
- Dunn, K., and Bremner, J. G. (2017). Investigating looking and social looking measures as an index of infant violation of expectation. *Dev. Sci.* 20, 1–6. doi: 10.1111/desc.12452
- Farroni, T., Massaccesi, S., Menon, E., and Johnson, M. H. (2007). Direct gaze modulates face recognition in young infants. *Cognition* 102, 396–404. doi: 10.1016/j.cognition.2006.01.007
- Feldman, R. (2012). Parent-infant synchrony: A biobehavioral model of mutual inferences in the formation of affiliative bonds: parent-infant synchrony. *Physiol. meas. emot. dev. perspect.* 77, 42–51. doi: 10.1111/j.1540-5834.2011.00660.x
- Field, A. (2005). *Discovering statistics using SPSS*. Thousand Oaks: Sage Publications.
- Finger, H., Goeke, C., Diekamp, D., Standvoß, K., and König, P. (2017). "LabVanced: a unified JavaScript framework for online studies [Conference paper]," in *International Conference on Computational Social Science IC2S2S, Cologne, (Germany)*.
- Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., et al. (2017). A collaborative approach to infant research: promoting reproducibility, best practices, and theory-building. *Infancy* 22, 421–435. doi: 10.1111/inf.12182
- Frank, M. C., Vul, E., and Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition* 110, 160–170. doi: 10.1016/j.cognition.2008.11.010
- Fraser, A., Gattas, S. U., Hurman, K., Robinson, M., Duta, M., and Scerif, G. (2021). Automated gaze direction scoring from videos collected online through conventional webcam. *PsyArXiv* 2021:5479.
- Gagné, N., and Franzen, L. (2021). How to run behavioural experiments online: best practice suggestions for cognitive psychology and neuroscience. *PsyArXiv* 2021:67. doi: 10.31234/osf.io/nt67j
- Gredebäck, G., Johnson, S., and von Hofsten, C. (2009). Eye tracking in infancy research. *Dev. Neuropsychol.* 35, 1–19. doi: 10.1080/87565640903325758
- Hannon, E. E., Schachner, A., and Nave-Blodgett, J. E. (2017). Babies know bad dancing when they see it: older but not younger infants discriminate between synchronous and asynchronous audiovisual musical displays. *J. Exp. Child Psychol.* 159, 159–174. doi: 10.1016/j.jecp.2017.01.006
- Hessels, R. S., and Hooge, I. T. C. (2019). Eye tracking in developmental cognitive neuroscience – The good, the bad and the ugly. *Dev. Cogn. Neurosci.* 40:100710. doi: 10.1016/j.dcn.2019.100710
- JASP Team (2021). *JASP (Version 0.14)* [Computer software]. Available online at: <https://jasp-stats.org/>
- Kominsky, J. F., Begus, K., Bass, I., Colantonio, J., Leonard, J. A., Mackey, A. P., et al. (2021a). Organizing the methodological toolbox: lessons learned from implementing developmental methods online. *Front. Psychol.* 12:702710. doi: 10.3389/fpsyg.2021.702710
- Kominsky, J. F., Gerstenberg, T., Pelz, M., Sheskin, M., Singmann, H., Schulz, L., et al. (2021b). The trajectory of counterfactual simulation in development. *Dev. Psychol.* 57, 253–268. doi: 10.1037/dev0001140
- Leshin, R., Leslie, S.-J., and Rhodes, M. (2020). Does it matter how we speak about social kinds? A large, pre-registered, online experimental study of how language shapes the development of essentialist beliefs. *Child Dev.* 2020:13527. doi: 10.1111/cdev.13527
- Lewkowicz, D. J. (1996). Perception of auditory-visual temporal synchrony in human infants. *J. Exp. Psychol.* 22, 1094–1106. doi: 10.1037/0096-1523.22.5.1094
- Libertus, K., and Violi, D. A. (2016). Sit to talk: relation between motor skills and language development in infancy. *Front. Psychol.* 7:475. doi: 10.3389/fpsyg.2016.00475
- Lingeman, J., Freeman, C., and Adolph, K. E. (2014). *Datavyu (Version 1.3.7)*. Datavyu.
- Lo, C. H., Mani, N., Kartushina, N., Mayor, J., and Hermes, J. (2021). e-Babylab: an open-source browser-based tool for unmoderated online developmental studies. *PsyArXiv* 2021:31234. doi: 10.31234/osf.io/u73sy
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during Covid-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Mathôt, S., and March, J. (2021). Conducting linguistic experiments online with OpenSesame and OSWeb. *PsyArXiv* 2021:31234. doi: 10.31234/osf.io/wnryc
- McClure, E. R., Chentsova-Dutton, Y. E., Holochwost, S. J., Parrott, W. G., and Barr, R. (2018). Look at that! Video chat and joint visual attention development among babies and toddlers. *Child Dev.* 89, 27–36. doi: 10.1111/cdev.12833
- Michel, C., Kayhan, E., Pauen, S., and Hoehl, S. (2021). Effects of reinforcement learning on gaze following of gaze and head direction in early infancy: an interactive eye-tracking study. *Child Dev.* 2021:13497. doi: 10.1111/cdev.13497
- Murray, L., and Trevarthen, C. (1986). The infant's role in mother-infant communications. *J. Child Lang.* 13, 15–29. doi: 10.1017/S030500090000271
- Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., and Hartley, C. A. (2020). Moving developmental research online: comparing in-lab and web-based studies of model-based reinforcement learning. *Collabra. Psychol.* 6:17213. doi: 10.1525/collabra.17213
- Nyström, M., Andersson, R., Holmqvist, K., and Van De Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behav. Res. Methods* 45, 272–288. doi: 10.3758/s13428-012-0247-4
- Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy* 22, 436–469. doi: 10.1111/inf.12186

- Oliver, B. R., and Pike, A. (2021). Introducing a novel online observation of parenting behavior: reliability and validation. *Parenting* 21, 168–183. doi: 10.1080/15295192.2019.1694838
- Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., and Hays, J. (2016). “WebGazer: scalable webcam eye tracking using user interactions,” in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, ed. S. Kambhampati (AAAI Press/International Joint Conferences on Artificial Intelligence), 3839–3845.
- Provati, J., Lemoine-Lardenois, C., Orriols, E., and Morange-Majoux, F. (2017). Do preterm infants perceive temporal synchrony? An analysis with the eye-tracking system. *Timing Time Percept.* 5, 190–209. doi: 10.1163/22134468-00002089
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Dev.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Richardson, E., Sheskin, M., and Keil, F. C. (2021). An illusion of self-sufficiency for learning about artifacts in scaffolded learners, but not observers. *Child Dev.* 2021:13506. doi: 10.1111/cdev.13506
- Ross-Sheehy, S., Reynolds, E., and Eschman, B. (2021). Unsupervised online assessment of visual working memory in 4- to 10-year-old children: array size influences capacity estimates and task performance. *Front. Psychol.* 12:692228. doi: 10.3389/fpsyg.2021.692228
- RStudio Team (2020). *RStudio: Integrated Development Environment for R* (Version 1.3.1093)[Computer Software]. Available online at: <https://www.rstudio.com/>
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (Part 2): assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/OPMI_a_00001
- Scott, K., and Schulz, L. (2017). Lookit (Part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Semmelmann, K., Hönekopp, A., and Weigelt, S. (2017). Looking tasks online: utilizing webcams to collect video data from home. *Front. Psychol.* 8:1582. doi: 10.3389/fpsyg.2017.01582
- Semmelmann, K., and Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behav. Res. Methods* 50, 451–465. doi: 10.3758/s13428-017-0913-7
- Sheskin, M., and Keil, F. (2018). TheChildLab.com A video chat platform for developmental research. *PsyArXiv* 2018:31234. doi: 10.31234/osf.io/rn7w5
- Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al. (2020). Online developmental science to foster innovation, access, and impact. *Trends Cogn. Sci.* 24, 675–678. doi: 10.1016/j.tics.2020.06.004
- Smith-Flores, A. S., Perez, J., Zhang, M. H., and Feigenson, L. (2021). Online measures of looking and learning in infancy. *PsyArXiv* 2021:31234. doi: 10.31234/osf.io/tdbnh
- Su, I.-A., and Ceci, S. (2021). “Zoom developmentalists”: home-based videoconferencing developmental research during COVID-19. *PsyArXiv* 2021:6. doi: 10.31234/osf.io/nvdy6
- The ManyBabies Consortium (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Adv. Methods Pract. Psychol. Sci.* 3, 24–52. doi: 10.1177/2515245919900809
- Tomar, S. (2006). Converting video formats with FFmpeg. *Linux J.* 2006:10.
- Tran, M., Cabral, L., Patel, R., and Cusack, R. (2017). Online recruitment and testing of infants with Mechanical Turk. *J. Exp. Child Psychol.* 156, 168–178. doi: 10.1016/j.jecp.2016.12.003
- Vallippan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., et al. (2020). Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nat. Comm.* 11:4553. doi: 10.1038/s41467-020-18360-5
- Venker, C. E., Pomper, R., Mahr, T., Edwards, J., Saffran, J., and Weismer, S. E. (2020). Comparing automatic eye tracking and manual gaze coding methods in young children with autism spectrum disorder. *Aut. Res.* 13, 271–283. doi: 10.1002/aur.2225
- Visser, I., Bergmann, C., Byers-Heinlein, K., Ben, R. D., Duch, W., Forbes, S. H., et al. (2021). Improving the generalizability of infant psychological research: The ManyBabies model. *PsyArXiv* 2021:8. doi: 10.31234/osf.io/8vwbfb
- Wass, S. V. (2016). “The use of eye-tracking with infants and children,” in *Practical Research with Children* 1st ed, eds J. Prior and J. Van Herwegen (Milton Park: Routledge), 24–45. doi: 10.4324/9781315676067
- Wass, S. V., Forssman, L., and Leppanen, J. (2014). Robustness and precision: how data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy* 19, 427–460. doi: 10.1111/inf.12055
- Winkler, I., Håden, G. P., Ladinig, O., Sziller, I., and Honing, H. (2009). Newborn infants detect the beat in music. *Proc. Natl. Acad. Sci.* 106, 2468–2471. doi: 10.1073/pnas.0809035106
- Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., and Xiao, J. (2015). TurkerGaze: crowdsourcing saliency with webcam based eye tracking. *ArXiv* 1504:06755v2.
- Yamamoto, H. W., Kawahara, M., and Tanaka, A. (2021). A web-based auditory and visual emotion perception task experiment with children and a comparison of lab data and web data. *Front. Psychol.* 12:702106. doi: 10.3389/fpsyg.2021.702106
- Zaadnoordijk, L., Buckler, H., Cusack, R., Tsuji, S., and Bergmann, C. (2021). A global perspective on testing infants online: introducing ManyBabies-AtHome. *PsyArXiv* 2021:5. doi: 10.31234/osf.io/cnwh5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Bánki, de Eccher, Falschlehner, Hoehl and Markova. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Remote Data Collection During a Pandemic: A New Approach for Assessing and Coding Multisensory Attention Skills in Infants and Young Children

Bret Eschman^{1*}, James Torrence Todd^{1,2}, Amin Sarafraz³, Elizabeth V. Edgar¹, Victoria Petrulla¹, Myriah McNew¹, William Gomez¹ and Lorraine E. Bahrick^{1,2}

¹ Department of Psychology, Florida International University, Miami, FL, United States, ² Center for Children and Families, Florida International University, Miami, FL, United States, ³ University of Miami Institute for Data Science and Computing, Miami, FL, United States

OPEN ACCESS

Edited by:

Rhodri Cusack,
Trinity College Institute of
Neuroscience, Ireland

Reviewed by:

Bennett I. Berthenthal,
Indiana University Bloomington,
United States

Jonathan F. Kominsky,
Harvard Graduate School of
Education, United States

*Correspondence:

Bret Eschman
beschman@fiu.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 27 June 2021

Accepted: 25 November 2021

Published: 21 January 2022

Citation:

Eschman B, Todd JT, Sarafraz A, Edgar EV, Petrulla V, McNew M, Gomez W and Bahrick LE (2022) Remote Data Collection During a Pandemic: A New Approach for Assessing and Coding Multisensory Attention Skills in Infants and Young Children. *Front. Psychol.* 12:731618. doi: 10.3389/fpsyg.2021.731618

In early 2020, in-person data collection dramatically slowed or was completely halted across the world as many labs were forced to close due to the COVID-19 pandemic. Developmental researchers who assess looking time (especially those who rely heavily on in-lab eye-tracking or live coding techniques) were forced to re-think their methods of data collection. While a variety of remote or online platforms are available for gathering behavioral data outside of the typical lab setting, few are specifically designed for collecting and processing looking time data in infants and young children. To address these challenges, our lab developed several novel approaches for continuing data collection and coding for a remotely administered audiovisual looking time protocol. *First*, we detail a comprehensive approach for successfully administering the Multisensory Attention Assessment Protocol (MAAP), developed by our lab to assess multisensory attention skills (MASKs; duration of looking, speed of shifting/disengaging, accuracy of audiovisual matching). The MAAP is administered from a distance (remotely) by using Zoom, Gorilla Experiment Builder, an internet connection, and a home computer. This new data collection approach has the advantage that participants can be tested in their homes. We discuss challenges and successes in implementing our approach for remote testing and data collection during an ongoing longitudinal project. *Second*, we detail an approach for estimating gaze direction and duration collected remotely from webcam recordings using a post processing toolkit (OpenFace) and demonstrate its effectiveness and precision. However, because OpenFace derives gaze estimates without translating them to an external frame of reference (i.e., the participant's screen), we developed a machine learning (ML) approach to overcome this limitation. Thus, *third*, we trained a ML algorithm [(artificial neural network (ANN))] to classify gaze estimates from OpenFace with respect to areas of interest (AOI) on the participant's screen (i.e., left, right, and center). We then demonstrate reliability between this approach and traditional coding

approaches (e.g., coding gaze live). The combination of OpenFace and ML will provide a method to automate the coding of looking time for data collected remotely. Finally, we outline a series of best practices for developmental researchers conducting remote data collection for looking time studies.

Keywords: gaze estimation, online data collection, remote data collection, looking time, Gorilla Experiment Builder, OpenFace, machine learning

INTRODUCTION

In early 2020, in-person participant testing and data collection dramatically slowed or was completely halted across the world as some labs were forced to close due to the COVID-19 pandemic. Developmental researchers who assess looking time (especially those who rely heavily on in-lab eye-tracking or live observer coding) were forced to re-think their methods of data collection. They could either analyze old data or they could attempt to adapt their data collection techniques to remote testing platforms—e.g., online data collection using an internet-connected computer in the child's home. During March of 2020, our lab was forced to close its doors to in-person participant testing in the middle of an extended longitudinal project. In an effort to continue data collection, we adapted many of our “in-lab” protocols and tasks to a format suitable for a remote setting. We found it relatively easy to convert parent questionnaires and assessments of children's language, social, and cognitive functioning, to this format. However, collecting looking time data for audiovisual tasks (i.e., tasks that track infant attention to multiple dynamic visual events in the presence of a soundtrack matching one of them) including the Multisensory Attention Assessment Protocol (MAAP; Bahrick et al., 2018a), posed significant challenges. For example, there are large individual differences in participants' home computer setups (e.g., differences in screen size, web camera quality, lighting, internet speed, etc.), making it difficult to use webcam-based eye-tracking techniques or to reliably code gaze in real-time. Further, because offline coding from videos (e.g., frame-by-frame) is time- and labor-intensive, we wanted to find a solution that might expedite the data coding process.

Fortunately, for those like us who opted to continue data collection during the pandemic, there are a variety of remote or online platforms (e.g., Amazon Mechanical Turk, Gorilla, Lookit, PyHab, etc.) that are specifically designed for gathering behavioral data outside of the typical lab setting. For example, Lookit has shown significant promise for remote data collection of looking time from infants and children (e.g., Scott and Schulz, 2017). It provides a secure, robust platform that can translate developmental methods to a computer-based home testing environment, affording greater accessibility to families both within and outside the university community. Similarly, Gorilla Experiment Builder is a promising tool for online data collection in adults and children, particularly for assessing executive functioning and working memory (Anwyl-Irvine et al., 2018; Ross-Sheehy et al., 2021), and it can also be used for collecting data from looking time tasks. While these platforms provide developmental researchers with legitimate options for online data collection, they have yet to be thoroughly vetted and

tested with infants and children in the home, integrated with audiovisual tasks, or integrated with reliable methods for gaze coding for audiovisual tasks.

For our purposes, we opted to use Gorilla Experiment Builder for the following reasons: (1) it provides excellent control of temporal parameters (e.g., trial onsets and offsets), (2) optimal video playback (e.g., little lagging, synchronous audio, and video playback), (3) an intuitive interface for building experiments so that they can be quickly deployed for remote data collection, and (4) can easily and efficiently be deployed in the home with minimal technology (e.g., experiments can be accessed through a web browser; Anwyl-Irvine et al., 2020a,b). In this paper, we detail our approach for collecting and coding looking time data remotely from our MAAP protocol—a three-screen (left, right, center displays) individual difference measure of three foundational attention skills (duration of looking, speed of shifting/disengaging, accuracy of audiovisual matching) to audiovisual social and non-social events (Bahrick et al., 2018a)—using widely available software and hardware available on home computers. We then describe our approach for scoring data from the MAAP that have been collected in the home, using a newly developed platform for estimating gaze behavior from video recordings (OpenFace), as well as our development of a machine learning (ML) model to translate the estimates provided by OpenFace into meaningful looking time data (i.e., left, right, and center displays). We end by discussing the implications this new approach for developmental researchers who are interested in collecting looking time data from infants and children remotely.

Traditional Methods for Coding Looking Time for Infants and Children

Developmental researchers who use looking time as an index of infant perception or cognition typically code it in one of three ways: using frame-by-frame coding, coding gaze in real time, or by using one of many different types of eye trackers. The general goal in using all of these methods is to estimate where the participant is looking on a screen, when they initiated the look (look onset), and how long they remain fixated on a particular location (look duration and offset). Typically, researchers define multiple areas of interest (AOIs) to demarcate locations on a screen displaying visual images that participants could view. For example, researchers have assessed looking time to the entire screen (e.g., Richards, 1987; Lewkowicz, 1988; Colombo et al., 1991), or locations corresponding to multiple images/events on a single screen (e.g., Hirsh-Pasek and Golinkoff, 1996; Bahrick et al., 2018b). While together, these looking time methods have generated a tremendous amount of information about the

development of attention, perception, and cognition, they require training coders (e.g., live coding and frame-by-frame coding), can be time consuming (e.g., frame-by-frame coding), and in some instances, cannot be adapted to an online setting (e.g., remote eye-tracking). The following is a brief overview of each of these methods for coding looking time and the problems that might arise when applied to coding data collected online.

Frame-by-Frame Coding

Frame-by-frame coding involves estimating gaze direction on each frame from a video recording (Fernald et al., 2008; Ross-Sheehy et al., 2015). Estimates of inter-rater reliability between human observers is typically very high (e.g., 85–95% agreement) but there appear to be limitations to the number of locations that can be reliably coded. This is potentially due to the relatively low spatial and temporal resolution when coding gaze direction from videos (Wass et al., 2013), making it difficult to code looking to more than two or three locations. These limitations are especially evident for data collected remotely as several of the environmental constraints that the lab setting affords (e.g., standardization of distance, position, and lighting) are absent. Further, this method of coding is extremely time consuming, and human coders can take up to 5 h to code 10 min of video (Wass et al., 2013). Due to this, frame-by-frame coding limits the amount of data that can be processed, and is typically used for shorter tasks (e.g., Jesse and Johnson, 2016). Further, human observers need to be trained and reliability must be established, both of which are also time consuming (Oakes, 2012).

Live Coding

Coding gaze in real time by trained observers is a widespread method for quantifying looking time. Observers, blind to the conditions of the study and unable to see the presentation of visual stimuli, estimate gaze direction and duration in real time while the participant views the stimuli (e.g., Lewkowicz, 1988; Bahrick et al., 2018a). This method is more time-efficient than frame-by-frame coding and requires little post processing of the data. In addition, if needed, observers can also code gaze offline from a video recording of the data collection. This approach has been used to estimate looking to a single location on a screen (e.g., Bahrick and Lickliter, 2000; Shaddy and Colombo, 2004; Altvater-Mackensen et al., 2016), looking to two locations (e.g., left, right; Bahrick and Watson, 1985; Bahrick, 1987; Casey and Richards, 1988), or looking to three locations (e.g., left, center, right; Bahrick et al., 2018a). However, for assessments administered remotely, coding data in real time is prohibitively difficult and offline coding of a video recorded via a webcam from a home computer can also be challenging. Specifically, without a clear frame of reference, it is difficult to judge where on the screen the participant is looking and how well the participant's looking is time-locked with the onsets and offsets of the visual stimulus events. Further, procedures should be used to ensure that observers are unaware of the experimental conditions (Oakes, 2012), which can also be difficult using online platforms.

Eye-Tracking

Eye-tracking has become an increasingly popular tool for examining looking time and has been developed and refined

over the last several decades (e.g., Hutton, 2019). Compared to methods using human observers, eye-tracking allows researchers to obtain gaze location objectively without the need to manually code the data (Hessels and Hooze, 2019) and features higher temporal and spatial resolution for gathering samples (Aslin, 2012; Wass et al., 2013). Gaze direction is determined by the reflection of infrared light sources on the eye(s) using information from the calibration process conducted prior to data collection. The calibration process stores information about the participant's pupil(s) and corneal reflection(s) for fixations at specified locations on the screen (Oakes, 2010, 2012). This allows for gaze to be measured in terms of X, Y coordinates for any location on the screen. However, infrared eye-tracking cannot be employed for remote data collection. Though webcam-based eye-trackers show some promise for remote data collection (Sammelmann and Weigelt, 2017), there has been little research into the feasibility of their use in collecting gaze data from infants and young children. For example, changes in participant's head position can lead to significant data loss, and calibration can be tedious and time consuming (increasing the likelihood of participant fatigue).

In sum, while the techniques reviewed above have provided a wealth of information derived from the looking behavior of infants and young children, they were not optimal (and in some instances, not possible) for our purposes of coding looking time from a protocol administered remotely. We sought to devise an approach in which looking time to a multi-screen audiovisual protocol (the MAAP; see section Our Audiovisual Task: The Multisensory Attention Assessment Protocol) could be coded accurately and efficiently across many participants. Thus, our approach does not supplant prior approaches, but rather provides researchers with a new tool for coding looking time data, one that is optimized for remote data collection. We detail our new approach in sections Objective 1: Data Collection at a Distance, Objective 2: Using OpenFace to Derive Gaze Estimates From Web-Cam Recordings, and Objective 3: Training a ML Model to Calculate Looking Time Data.

Online Data Collection

In addition to the challenges of coding looking time in an online setting, there are a number of challenges specific to online data collection. For example, a unique problem with online (internet-based) testing is its reliance on participants' home computer hardware and software. In the lab, researchers develop and refine their lab computer, stimulus software, and hardware for data collection. More important, they can be sure that all participants are tested using the same system. For online testing, the opposite is true: participants use different computers (desktop, laptop, tablet, or even phone), as well as different operating systems and web browsers. Because of this, ensuring uniform standards for data collection is extremely difficult. While all of the unique combinations of hardware and software are not equal, some home computer set-ups outperform others (Anwyl-Irvine et al., 2020a,b). By limiting the number of platforms (described below), designing experiments that require minimum amounts of technology, and providing the participants/caregivers with explicit detail on how to set up their home computer, we

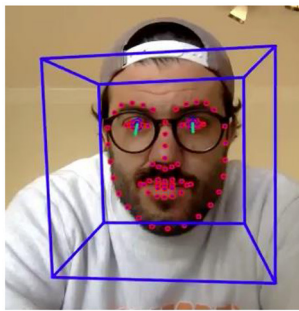


FIGURE 1 | Example of OpenFace output including facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation. These metrics are computed for each frame of the video. Depicted is a single frame from the video.

can more closely recreate the lab setting using remote testing in the home.

Post-processing of Looking Time Data

Recent advances in the offline post processing of looking time data provide researchers with a viable option for scoring data collected remotely (as well as in the lab) from video recordings. In fact, there are a wide variety of open-source tools and commercial systems available for eye-gaze estimation (Wood and Bulling, 2014; Baltrusaitis et al., 2016; Park et al., 2018; Chouinard et al., 2019) and webcam-based eye tracking solutions (e.g., <https://github.com/stepacool/Eye-Tracker> and <https://webgazer.cs.brown.edu/>). Further, Chouinard et al. (2019) used an automatic face analysis tool (Amazon Rekognition) to automate infant preferential looking coding from video data collected online.

We found that one tool in particular (OpenFace; <https://github.com/TadasBaltrusaitis/OpenFace#eye-gaze-tracking>) seems to be well-suited to address our specific needs of coding looking time data collected in a remote setting (for our particular three-screen audio visual task) with infants and young children. We chose to use OpenFace for eye-gaze estimation in our current project for the following reasons. OpenFace is an open source, post processing, gaze estimation tool (the code is freely available for academic purposes). It is the first toolkit with available source code capable of facial landmark detection, head pose estimation, facial action unit recognition, and, most importantly, eye-gaze estimation (see **Figure 1**). Further, this tool can estimate these parameters from video recordings of a participant's face.

Although there has been an increased interest in automatic gaze estimation analysis and understanding, there has been little application of these techniques to infants and young children (but see Chouinard et al., 2019). Further little is known about how these techniques may be integrated with online data collection—using technology owned by most families (e.g., laptops, basic internet, standard webcams).

The Current Project

The current project has three main objectives. *First*, we lay out a detailed approach for successfully collecting looking time data

from a distance. Using only Zoom, an online experiment builder (Gorilla), an internet connection, and equipment typically found in the home (e.g., laptop, webcam, speakers), we have developed a novel, successful method for administering an audiovisual looking time task (the MAAP) remotely. *Second*, we detail an approach for estimating gaze direction and duration collected remotely from webcam recordings using OpenFace. While OpenFace provides us with estimates of gaze direction/vectors, these estimates are meaningless in the absence of an external frame of reference (specific location on the participants screen). *Third*, to overcome this challenge, we developed a novel ML approach for training an algorithm (neural network) to classify gaze direction/vectors into traditional looking time data (e.g., total looking to left, right, and center displays) by relating gaze directions from OpenFace to an external frame of reference (locations on the participant's screen). To assess accuracy of these looking time estimates, we assess reliability between these looking time estimates using OpenFace/ML and the same data that were previously coded live (traditional approach) from a longitudinal study conducted in our lab using the MAAP (Bahrick et al., 2018b). The data that were coded live serve as the baseline and proof of concept for using OpenFace/ML to code looking time data. We then discuss applying this novel approach to data collected in participants' homes from webcam recordings. We provide the ML model with a series of "known locations" (attention getting stimuli), to define looks to left, right, and center. Further, we provide a set of guidelines for how to implement looking time measures in the home, with minimal software and equipment. This novel approach is designed to address the immediate need of continuing data collection during a pandemic (or any lab shutdown) by combining a variety of methods into a single framework. In addition to serving this immediate purpose, it is our hope that this method can be developed further to offer future researchers a viable method for collecting meaningful data remotely.

Our Audiovisual Task: The Multisensory Attention Assessment Protocol

We demonstrate the effectiveness of this approach to online data collection and the OpenFace post processing method using data collected in our lab from the MAAP (Bahrick et al., 2018a). The MAAP is a fine-grained measure of individual differences in attention to dynamic, audiovisual social, and non-social events, appropriate for infants and young children. The MAAP assesses three multisensory attention skills (MASKS; duration of looking, speed of shifting/disengaging, accuracy of audiovisual matching) using 24 short trials (to provide stable means), presents blocks of both social and non-social events, and indexes the cost of competing stimulation from a visual distractor event on each of these skills. Trials of the MAAP consist of a 3-s dynamic, silent central event (morphing geometric shapes) followed by two 12-s side-by-side lateral events of women speaking (social events) or objects impacting a surface in an erratic pattern (non-social events). One of the lateral events is synchronous with its natural soundtrack, and the other lateral event is asynchronous with the soundtrack. For an example video, visit <https://nyu.databrary.org>.

org/volume/326. Performance on the MAAP predicts language outcomes in typically developing infants and children (Bahrick et al., 2018a; Edgar et al., Under review¹), and predicts language and symptomatology in children with autism (Todd and Bahrick, Under review)². Also, unlike prior research using static images or silent events, by presenting audiovisual events on three displays, and presenting both social and non-social events in the presence of an irrelevant visual distractor, the MAAP better reflects the natural, multisensory learning environment of the child. Further, the MAAP requires no verbal responses or verbal instructions to the child, and is thus able to provide a common measure for assessing development across infancy and early childhood.

OBJECTIVE 1: DATA COLLECTION AT A DISTANCE

We adapted two well-established remote data collection platforms (Zoom and Gorilla) for use with technology that is commonly found in the home. Zoom provides videoconferencing and online chat services through a cloud-based peer-to-peer software platform (<https://zoom.us/>). This platform provides a stable environment for real time face-to-face communication, including live interaction allowing the experimenter to provide instructions and guidance, as well as the opportunity for the participants to ask questions or provide feedback. Sessions can be recorded for later behavioral coding. Gorilla (www.gorilla.sc) is an online experiment builder whose aim is to enable researchers to conduct online experiments (regardless of programming and networking knowledge). It provides access to web-based experiments and reduces the risk of introducing noise (e.g., misuse of browser-based technology) in data (Anwyl-Irvine et al., 2018). Combined, these two platforms can be used to conduct looking time tasks in the home with acceptable precision and accuracy for temporal parameters (Anwyl-Irvine et al., 2020a,b).

Programming and Presenting the Task

After programming a version of the MAAP that ensured sufficient audio-visual synchrony (see **Supplementary Material**, section 8.1, for details), we used Gorilla Experiment Builder to present it to the participants in their own homes with minimal technical requirements. In addition to providing a platform for programming experiments, Gorilla Experiment Builder provides a straightforward way to present stimuli to participants on their own computer. Gorilla packages the task in a link that can be shared and displayed using a web browser. After the participant clicks on the link, the task will be displayed like a standard web

page. Participants are not required to download anything; they simply click, and the program is launched.

Administering the MAAP Remotely

Although there are methods available to collect looking time data in Gorilla (e.g., webcam based eye-tracking), we found them somewhat difficult to work with and, importantly, the integration of the webcam eye-tracking software with the videos introduced some noise (e.g., lagging videos, asynchrony of video, and audio soundtrack), into the presentation of the MAAP. Because the MAAP depends on ensuring that the audio track aligns with the video, it was imperative that we were able to record looks while maintaining excellent audio-visual precision between the video and the audio track. We found that a combination of Gorilla (not including their webcam-based eye tracking feature) and Zoom provided us with the level of precision that we required. Specifically, our preliminary tests indicated that Gorilla playback through the Zoom share screen function achieved a sufficient level of precision of video and audio playback to allow us to code looking time data from the MAAP. While this sounds simple enough, there are a number of specific settings that the experimenter needed to enable in order to maintain the level of precision needed for this task (see **Supplementary Material**, section 8.2, for details).

The Role of the Caregiver

In addition to the technical requirements on the experimenter's end (see **Supplementary Material**, section 8.3.1), we also found that we needed to ensure that the parent/caregiver had the necessary technology to participate. Because we wanted to maximize time with the child and reduce demands on their attention, we found it helpful to have a “pre-session” with the parent/caregiver to familiarize them with the software (see **Supplementary Material**, section 8.4, for a description of this pre-session). In this pre-session we assessed their level of comfort with Zoom and familiarized them with its features, if necessary. Next we assessed what kind of computer they were using. While most desktop/laptops were compatible, and Gorilla demonstrates similar performance across both Macs and PCs (Anwyl-Irvine et al., 2020a,b), we found that it was not possible for parents to use tablets or computers that did not have a webcam at the top center of the screen. This is because we needed consistency for our post processing method (e.g., same camera location and distance of the child from the camera). Next, we tested the participant's internet speed. Using <https://www.speedtest.net/>, we recorded the participant's download speed and their ping rate. We found that as long as their download speed was >50 mbps, and their ping was <25 ms, there were no issues in terms of lagging or asynchronous presentation of the MAAP (see **Supplementary Material**, section 8.3.2, for more detail).

We also found it imperative to discuss with the parent/caregiver the importance of their role in testing and data collection. Specifically, we wanted to emphasize that the parent was not a passive viewer of the data collection, but rather an “active at-home experimenter,” working alongside the experimenters from our lab. By giving parents this title and providing specific instructions for how to best approach the data

¹Edgar, E. V., Todd, J. T., and Bahrick, L. E. (Under review). *Intersensory Matching of Faces and Voices in Infancy Predicts Language Outcomes in Young Children*. [Manuscript submitted for publication]. Department of Psychology, Florida International University.

²Todd, J. T., and Bahrick, L. E. (Under review). *Individual Differences in Multisensory Attention Skills in Children with Autism Spectrum Disorder Predict Language and Symptom Severity: Evidence from the Multisensory Attention Assessment Protocol (MAAP)* [Manuscript submitted for publication]. Department of Psychology, Florida International University.

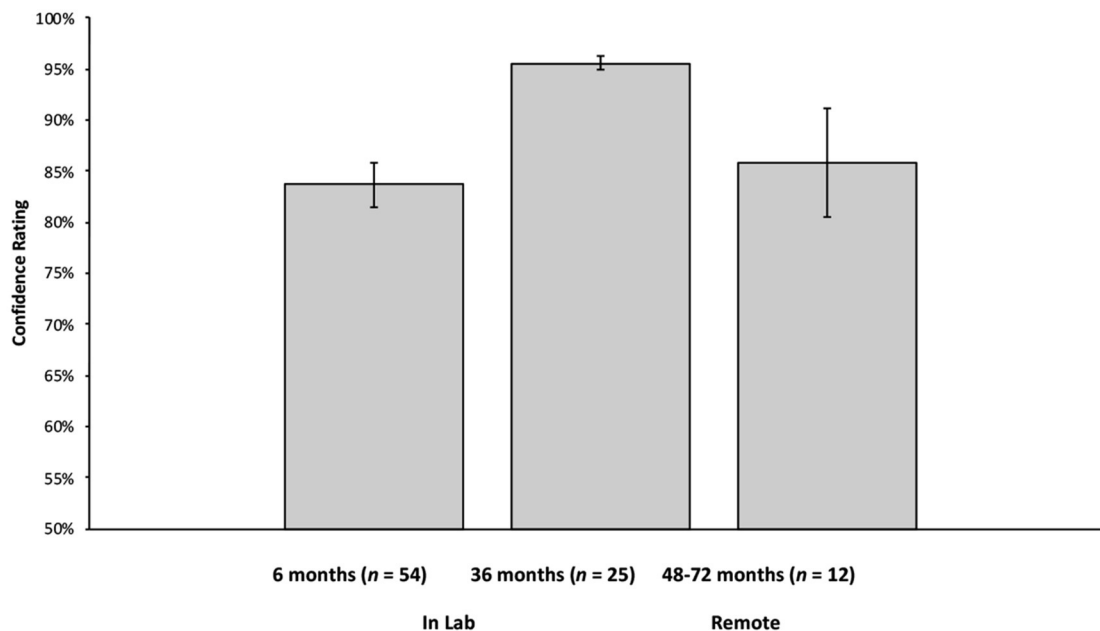


FIGURE 2 | Mean confidence ratings for 6-month-old infants (tested in the lab), 36-month-old children (tested in the lab), and 48–72-month-old children (tested remotely). Error bars reflect standard errors of the means. OpenFace provides a confidence rating for every frame of a video (using all face, head position, and gaze direction landmarks). The confidence rating is a measure of how well OpenFace can identify all of these components (averaged across all frames for each participant).

collection process, we found that parents were more engaged with the data collection process. We started by stressing that they needed to be present for the duration of the session and that they must be ready to help at any time. Helping included: setting up the camera angles, providing technical support (ensuring the tasks opened and were displayed correctly), and keeping the child engaged with the task. We also emphasized that they should not interfere with the data collection and that only the actual experimenter (who was present during task administration to give feedback and instructions) should give feedback to the child. Children were given no instructions about where to look. They were told that they would “watch some videos of ladies talking and objects moving,” that they needed to “sit nice and still,” and that after the video was done, they could “play a *really* fun game.” All of the information in the pre-session was recorded.

OBJECTIVE 2: USING OPENFACE TO DERIVE GAZE ESTIMATES FROM WEB-CAM RECORDINGS

While the combination of research tools mentioned in section Objective 1: Data Collection at a Distance provides a promising and exciting method for remote data collection, it does not address the issues of quantifying and processing looking time data. Therefore, the video recordings of the participants completing the task require additional post-processing through OpenFace to estimate gaze direction/vectors (for a summary of OpenFace features, see section Post Processing of Looking

Time Data). To demonstrate the effectiveness of using OpenFace to estimate gaze direction in infants and young children from a video recording, we first evaluate how well OpenFace can identify the face, head position, and eye gaze direction of the participants. OpenFace provides a confidence rating for every frame of a video (using multiple face, head position, and gaze direction landmarks). The confidence rating is a measure of OpenFace’s accuracy in identifying all components. They range from 0 to 100% (higher is better). Frames in which the participant is looking away from the camera or occluding their face would receive a low confidence rating. It should be noted that we did not initially set any criteria for data inclusion. As such, the values provided by OpenFace are raw and unfiltered. Below, we describe the confidence ratings for data collected in the lab (as a proof of concept) and then extended this approach to data that were collected remotely.

OpenFace Confidence Ratings for Data Collected In-lab

Six- and 36-month-olds received the MAAP as a part of an ongoing longitudinal study. The longitudinal study, entitled “[blinded],” received IRB approval from the Social and Behavioral Review Board of [blinded] (IRB-13-0448-CR06). Video recordings of these sessions were processed by OpenFace as a proof of concept for this approach. Video data (including videos processed by OpenFace) are stored on a secure university server, can only be accessed by trained lab personnel, and can only be identified via a master key, which is kept in a separate, physical location. A summary of the confidence ratings can be

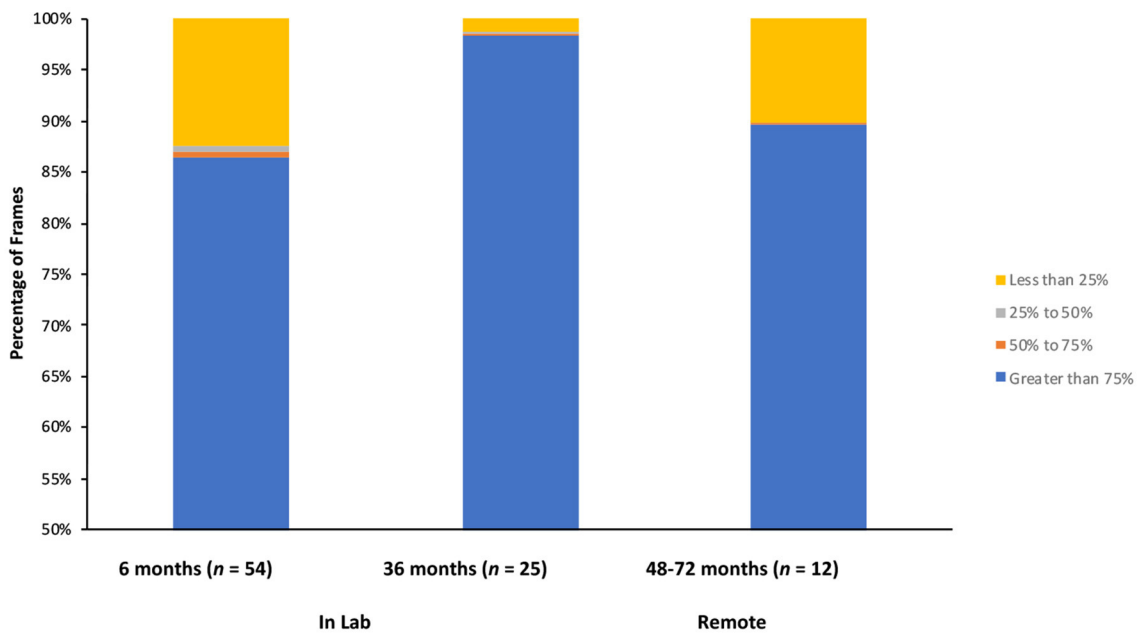


FIGURE 3 | Distribution of confidence ratings: percentage of frames with confidence ratings of 75% or higher (in blue), 50–75% (orange), 25–50% (gray), or <25% (yellow) for 6-month-old infants (tested in the lab), 36-month-old children (tested in the lab), and 48–72-month-old children (tested remotely). For each participant, we calculated the percentage of frames in each quartile, and then averaged across participants to get these numbers.

found in **Figure 2** and in general, were quite accurate. Averaged across all frames, 6-month-old infants ($n = 54$; tested in the lab) had an overall confidence rating of 83.68% ($SD = 16.11\%$), 36-month-olds ($n = 26$; tested in the lab) had a confidence rating of 95.55% ($SD = 3.37\%$). Further, inspection of the distribution of confidence ratings revealed that, at 6 months, 86.35% of frames analyzed by OpenFace had a confidence rating of 75% or higher, and at 36 months, 98.27% had a confidence rating of 75% or higher (see **Figure 3**).

OpenFace Confidence Ratings for Data Collected Remotely

Video recordings of 48–72-month old children who participated in the MAAP remotely were also processed via OpenFace. Confidence ratings for children tested in the home ($n = 12$) were also quite accurate (see **Figure 2** for a summary). Averaged across all frames, they had an overall confidence rating of 85.88% ($SD = 18.51\%$). Further, inspection of the distribution of confidence ratings revealed that, 89.97% of frames analyzed by OpenFace had a confidence rating 75% or higher (see **Figure 3**).

Challenges of Processing the Data

Because standard methods for coding looking time (see section Traditional Methods for Coding Looking Time for Infants and Children) are either too time consuming or impossible to use in a remote setting, we opted to automate the coding process, using OpenFace. OpenFace derives X, Y, Z (3D) coordinates of gaze direction and facial landmarks from the image of the participant's face on each video frame, but without translating these coordinates to an external frame of reference (i.e., locations

on the participant's computer screen). This means that we don't know precisely where the participant is looking on the screen using the OpenFace output alone, given variability across participants in properties of the camera lens and visual angles. As a result, we developed a ML approach to overcome this lack of information about gaze direction with respect to the external frame of reference. To do this, we trained a ML model to classify gaze direction with respect to specific AOI on the screen, based on vector information provided by OpenFace. By training an ML algorithm in this manner, we can use the estimates provided by OpenFace to predict individual look directions and durations (traditional measures for looking time studies) for each participant (in our case, looking to the left, center, and right displays in the MAAP).

Processing the Data Using Machine Learning and OpenFace

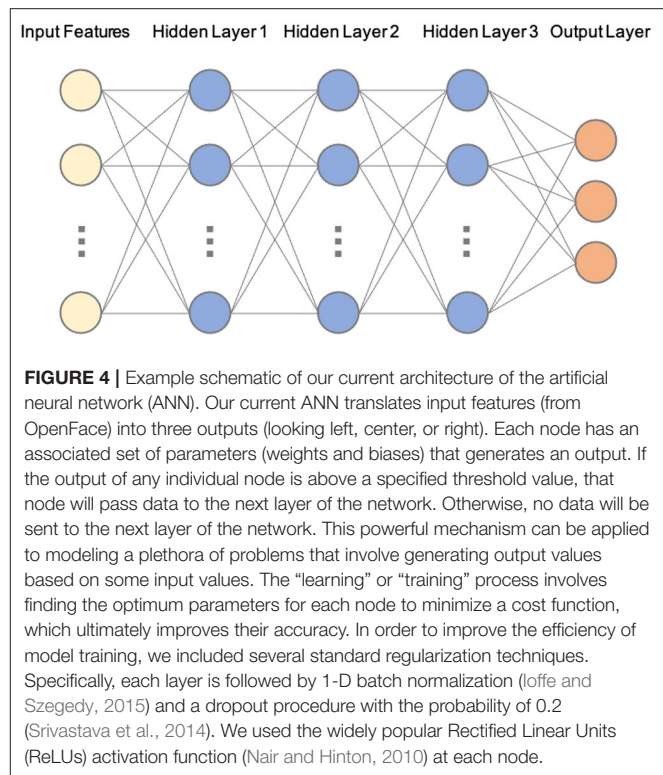
Machine learning is a data-driven approach for classifying patterns of relations between two or more variables typically from a large subset of a dataset (e.g., 50% of trials) in order to predict patterns of relations between these same variables in another subset of the data (e.g., remaining 50% of trials). ML algorithms accomplish this task by leveraging large amounts of data and computational power, and have been used in many disciplines (e.g., healthcare, autonomous driving, product recommendations). Artificial Neural Networks (ANNs) and their more advanced variants (Deep Neural Networks) are a widely used subset of ML approaches inspired by and based on biological neural networks. They are typically comprised of multiple

connected node layers that translate a set of inputs (e.g., the X, Y, Z coordinates provided by OpenFace) into outputs (looks to left, center, and right displays on the MAAP). The “learning” or “training” process in ANNs is a powerful learning mechanism that can ultimately improve the accuracy of the network when presented with new data. This approach is similar to multi-voxel pattern analysis (MVPA) used to predict patterns of relations of fMRI and fNIRS data (e.g., Norman et al., 2006; Emberson et al., 2017). For example, MVPA can be used to predict neural activity in a single infant based on data patterns classified across the rest of the infants in a sample, or can predict patterns in one or more trials from patterns classified across the rest of the trials of a single infant (e.g., Emberson et al., 2017).

While there have been previous successful examples of using a combination of eye tracking and ML to estimate gaze localizations (e.g., George and Routray, 2016; Akinyelu and Blignaut, 2020; and for a review, see Klaib et al., 2021), our approach was developed specifically for our three-screen video based protocol to be used with infants and children with data collected remotely and thus complements these previous approaches. Our goal was to develop a ML model that could be easily used and understood by individuals with little or no prior ML modeling experience. As such, we chose to use a multi-layer ANN as our current ML algorithm (see **Figure 4**). After preliminary testing, we adopted a network consisting of four fully connected (three hidden) layers which was the minimum architecture effective for our specific needs. Our first model included one hidden layer. In subsequent model development, we tried a variety of layers and nodes (i.e., hyper parameters), ultimately settling on the three layers in the current model, which demonstrated excellent agreement with trained human coders. It should also be noted that we intentionally chose a simple architecture to evaluate the effectiveness of ML approaches in solving our unique problem. This simple yet effective neural network can also serve as a baseline for comparing the performance of more advanced deep learning techniques.

Here, we demonstrate the feasibility of using a simple ANN approach within individual infants, classifying patterns of relations between two variables (data coded by live observers and X, Y, Z coordinates provided by OpenFace) in a subset of the data (50% of video frames) in order to predict relations between these same variables in separate subset of the dataset (e.g., 50% of frames) from the same participant. Our goal is to design an algorithm that can use the information from a series of single images (one for each frame of our video recording) that is extracted when processing the video through OpenFace. Specifically, we use the following information from OpenFace as input to the model:

- Eye gaze direction vector and their average for both eyes
- Location of 2D and 3D eye region landmarks
- Pose estimates: location and rotation of the head with respect to camera
- Face Landmarks locations in 2D and 3D space
- Rigid and non-rigid shape parameters
- Facial Action Units.



More information about each one of the above inputs is available at <https://github.com/TadasBaltrusaitis/OpenFace/wiki/Output-Format>. The “learning” or “training” process in ML involves finding the optimum combination of parameters that can most efficiently predict an outcome (e.g., gaze direction).

OBJECTIVE 3: TRAINING A ML MODEL TO CALCULATE LOOKING TIME DATA

In order to calculate traditional looking time measures from gaze estimation vectors from OpenFace, we trained a ML model to classify gaze estimation vectors to looks to the left, right, and center displays of the screen (AOIs) during the MAAP protocol. We then compared ML estimates to estimates provided by human observers who coded data in the lab. For the data that were collected in the lab, look directions (to left, right, and center displays on MAAP trials) were coded live by trained observers according to standard procedures used in infant studies (e.g., Casey and Richards, 1988; Shaddy and Colombo, 2004; Bahrick et al., 2018a). Specifically, looking time and direction were coded by a primary and a secondary observer during task administration. Observers, hidden behind a black curtain, viewed the child through a front facing camera (SONY FDR-AX33) hidden above the widescreen monitor. Observers were blind to condition, and they coded infant fixations to the left, center, and right sides of the screen in real-time using a game pad. Button presses were fed into a custom computer program

that calculated individual looking time to left center and right. Interobserver reliability was assessed by having the secondary observer record the looking for 66% of the participants ($n = 36$) at 6 months and 40% ($n = 10$) of the participants at 36 months. We assessed interobserver reliability by calculating the absolute difference between estimates of the two coders. To the extent that measures are free of random error (i.e., reliable), scores from each observer should be comparable (difference close to zero). This method is superior to correlational approaches for assessing inter-observer reliability, which are subject to artifacts (Goodwin and Leech, 2006; Jaccard and Becker, 2009). Inspection of the median absolute differences relative to the range of possible scores for each measure indicates little difference between the scores of the two observers and thus excellent reliabilities (differences were close to 0: from 0.009 to 0.053).

In order to train the model, we used looking time data that were coded live from the primary observer and then translated from individual look durations into frame-by-frame data. For this initial stage of ML training, we only used a subset of the total number of the 24 MAAP trials—a block of six social trials. All participants had a minimum of five out of six trials. From these six trials, for each participant we randomly selected 50% of the 3,270 total frames for the ML training set and used the remaining 50% of frames for the testing set (to assess agreement between ML estimates and the estimates of the trained coder). For our next steps, we will use 50% of the entire 24 trials for the training set and the remaining 50% for the testing set to assess agreement. Our protocol was designed such that the size, location, and trial duration of left, center, and right displays are identical across social and non-social conditions. Thus, we anticipate strong agreement between ML estimates and a trained coder across all 24 trials (social and non-social) on the MAAP. Importantly, ML algorithms rely on multiple training iterations to learn and improve their accuracy, meaning the predictions should improve each time these training iterations are completed.

For the data that were collected online, because live coding was not possible, we assessed agreement between ML estimates of the child's looking behavior and the known locations of attention getting stimuli. We recorded videos of the participants watching attention getting stimuli and then used OpenFace to estimate gaze locations to the screen. The attention getting stimuli were presented in the middle of each of the AOIs (i.e., left, center, and right). They were presented one at a time, starting in the center, then, left, then back to center, and then to the right. This sequence was then repeated. Attention getting stimuli were presented for 1,500 ms each. Unlike previous attempts to localize gaze using ML (e.g., George and Routray, 2016), participants were not explicitly instructed to fixate each point. This was in part because our sample consisted of young children. Further, the attention getting stimuli were designed to be highly salient so that the children would fixate them. In addition to appearing rapidly, during the 1,500 ms presentation time, each point changed color, grew and then subsequently shrunk in size, and was accompanied by a series of salient sounds. This provided several known locations on the screen for each participant, for short periods

of looking before the task started, serving as an external frame of reference. Importantly, when these attention getting stimuli were presented on the screen, they were the only thing visible. Therefore, we can assume that if the participant was looking at the screen, they were fixating each point. This allowed us to provide the model with a set of parameters for each of the three looking locations. Just like the approach described above for the data collected in the lab, 50% of data for each participant was used for training and validation of the ML algorithm and the rest was used for testing the performance of the algorithm.

After demonstrating the effectiveness of using OpenFace to quantify looking time in infants and young children by using the face, head position, and eye gaze direction of the participants (section OpenFace Confidence Ratings for Data Collected In-lab), we then compared the gaze estimate (provided by OpenFace) to a known location. To evaluate these outputs, we computed a percent agreement rate, or the number of frames where the OpenFace output provided the same estimate (e.g., left, center, right) as the live coder or as the known location (attention getting stimuli), divided by the total frames. For the data that were collected in the lab, we used data that were coded live (during data collection), and for the data collected remotely, this consisted of using attention getting stimuli (to provide a ground truth).

Predicting Data Collected In-lab

For data collected in the lab, agreement between predictions of the ML model and live coders was calculated on the 50% of frames not used for training (i.e., the testing set). Assuming an equal distribution of looks to left, center, and right, without model training, the initial predictions of the model should be at chance (33%). However, training the model should result in dramatic improvement in agreement. For example, following model training, for 6-month-old infants ($n = 54$), the ML model had an average agreement rate of 89.9% with the live coders ($SD = 6.75\%$). Further, this result was not driven by any one location on the screen as agreement scores were 82, 88, and 87% to the left, center, right, respectively. There were no significant differences in agreement among left, center, and right displays ($ps > 0.199$). It is important to note that because the amount of looking to center, right and left locations differed, the average agreement rate computed was a weighted average and thus does not correspond precisely to the mean derived by averaging across the scores for the three locations (left, right, and center displays). Similarly, for 36-month-olds ($n = 25$), the ML model had an average agreement rate of 85.83% with the live coders ($SD = 5.85$), with 85, 86, 80% agreement to the left, center, and right (with marginally greater agreement to the center display, 86%, than right display, 80%, $p = 0.092$). Two 6-month-old infants had very poor agreement to individual locations. Participant A had only 31% agreement to the center location and Participant B had only 0.05% agreement to the left location. Further inspection revealed that OpenFace appeared to have difficulties identifying all of the necessary input components due to the fact that Participant A's hand was obscuring the face for large parts of



FIGURE 5 | Participant A (in-lab). Example of a 6-month-old infant that OpenFace may have trouble identifying all components (facial landmark, head pose estimation, facial action unit recognition, and eye-gaze estimation) due to obstruction (hand) of the child's face.



FIGURE 6 | Participant B (in-lab). Example of a 6-month-old infant that OpenFace may have trouble identifying all components due to the fact that the video did not capture all of the child's face.

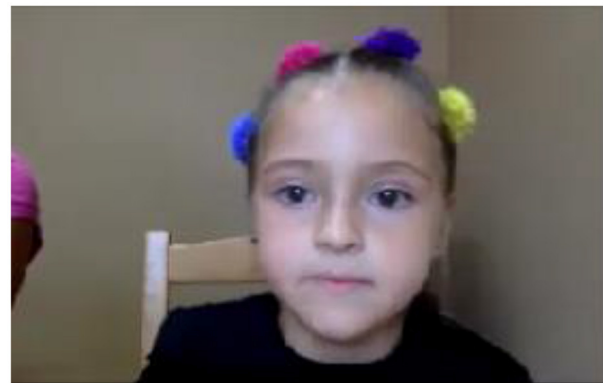


FIGURE 7 | Exemplary video for OpenFace.

TABLE 1 | In lab ML agreement without participants A and B.

| Age (months) | <i>n</i> | <i>M</i> (%) | <i>SD</i> (%) | Range (%) |
|--------------|----------|--------------|---------------|--------------|
| 6 | 52 | 90.18 | 6.63 | 73.51–99% |
| 36 | 25 | 85.83 | 5.85 | 72.57–95.12% |
| Total | 77 | 88.77 | 6.67 | 72.51–99% |

One participant had an average agreement score of 7.85%. Again, this appeared to be due to the fact that the child's face was not entirely in the frame while the attention getting stimuli were being presented (see **Figure 8**). When that participant was removed (Participant C), average agreement between the ML model and the attention getting stimuli improved to 90.7% ($SD = 7.06\%$; **Table 2**) and to 89, 91, and 88% for looking to the left, center, and right. Therefore, when the full face is within view during calibration, the ML model seems to have excellent potential for translating the OpenFace output into gaze locations to specific AOIs on the screen for individual participants. Thus, again when using OpenFace, an important inclusion criterion should be that participant faces are fully visible.

Demographic Differences

Because facial recognition software can sometimes be biased toward or perform better with members from the ethnic or racial group that developed it (e.g., Mehrabi et al., 2019), we explored the ability of OpenFace to identify and predict gaze across individuals who differed in gender, ethnicity, and race. Importantly, the ability of OpenFace to identify facial landmarks (i.e., confidence ratings) did not significantly differ as a function of gender ($p = 0.761$), Race ($p = 0.227$), or Ethnicity ($p = 0.170$). Additionally, percent agreement between the ML model and estimates from in-lab and remotely administered experiments did not significantly differ as a function of gender ($p = 0.219$), Race ($p = 0.189$), or Ethnicity ($p = 0.520$). See **Tables 3–5** for means and standard deviations.

the task (see **Figure 5**) and Participant B's face was not entirely in the video for large sections (see **Figure 6**). An optimal set up can be seen in **Figure 7**. With these two individuals removed from the dataset, average percent agreement for the 6-month-old infants improved to 90.18% ($SD = 6.63\%$; **Table 1**) and looking the left, center, and right improves to 84, 89, and 87% agreement, respectively. Therefore, in developing inclusion criteria for this and future attempts using OpenFace, one should ensure that participant faces are fully visible.

Predicting Data Collected Remotely

Because the data collected remotely could not be coded live, the ML model used attention getting stimuli (at the beginning of each block) as a frame of reference, similar to the procedure used by most remote eye-trackers.

Following training, the average percent agreement for data collected remotely (48-, 60-, and 72-month-old children; $n = 12$) between the ML model and the attention getting stimuli was 85.63% ($SD = 24.76\%$) with 82, 84, and 83% agreement to the left, center, and right locations. There were no significant differences in agreement among left, center, and right displays ($ps > 0.552$).

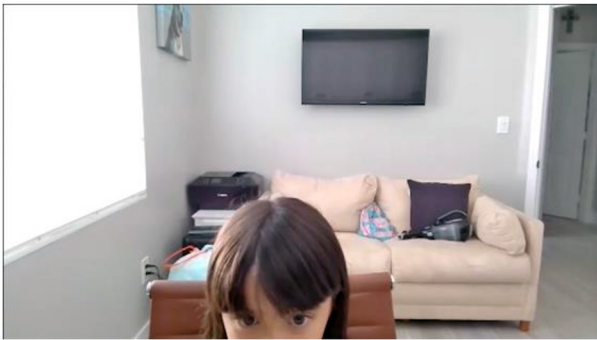


FIGURE 8 | Participant C (at-home). Example of a 72-month-old child that OpenFace may have trouble identifying all components due to the fact that the video did not capture all of the child's face.

TABLE 2 | Online ML agreement without participant C.

| Age (months) | <i>n</i> | <i>M</i> (%) | <i>SD</i> (%) | Range (%) |
|--------------|----------|--------------|---------------|-------------|
| 48 | 2 | 89.73 | 8.87 | 83.45–96 |
| 60 | 3 | 95.14 | 2.68 | 92.07–97.06 |
| 72 | 6 | 88.8 | 8.03 | 76.43–96 |
| Total | 11 | 90.7 | 7.06 | 76.43–97.06 |

TABLE 3 | Confidence rating (OpenFace) and ML agreement as a function of gender.

| | Gender | <i>N</i> | <i>M</i> (%) | <i>SD</i> (%) |
|-------------------|--------|----------|--------------|---------------|
| Confidence rating | Male | 45 | 87 | 15.90 |
| | Female | 37 | 86 | 15.45 |
| ML agreement | Male | 45 | 89 | 6.94 |
| | Female | 37 | 86 | 14.65 |

OpenFace provides a confidence rating for every frame of a video (using all face, head position, and gaze direction landmarks). The confidence rating is a measure of how well OpenFace can identify all of these components (averaged across all frames for each participant).

DISCUSSION

In this article, we have described our novel and successful method of data collection during the COVID-19 pandemic. While there are inherent challenges to testing remotely, and even more challenges when testing children, we found that with the proper attention to detail, very good quality data can be collected. Using a combination of Zoom and Gorilla Experiment Builder, looking time tasks can be programmed and used in the home easily and efficiently. Importantly, Gorilla afforded us with the level of audio-visual precision that was necessary for our looking time task. We also found it important to provide the caregiver with explicit instructions during a pre-session for how to help with data collection and serve as the “at-home experimenter” and facilitate testing without interfering. Once the data were collected, we used OpenFace (an open source gaze estimation tool) and a ML model to process the data. We developed a ML approach that was intentionally simple (compared to other deep

TABLE 4 | Confidence rating (OpenFace) and ML agreement as a function of race.

| | Race | <i>N</i> | <i>M</i> (%) | <i>SD</i> (%) |
|-------------------|------------------|----------|--------------|---------------|
| Confidence rating | African American | 11 | 77 | 24.69 |
| | White | 57 | 87 | 14.33 |
| | Other | 2 | 80 | 8.88 |
| | More than 1 race | 6 | 91 | 11.04 |
| | DNA | 6 | 94 | 3.11 |
| ML agreement | African American | 11 | 93 | 4.21 |
| | White | 57 | 85 | 12.46 |
| | Other | 2 | 87 | 8.71 |
| | More than 1 race | 6 | 91 | 5.74 |
| | DNA | 6 | 93 | 11.13 |

TABLE 5 | Confidence rating (OpenFace) and ML agreement as a function of ethnicity.

| | Ethnicity | <i>N</i> | <i>M</i> (%) | <i>SD</i> (%) |
|-------------------|------------------------|----------|--------------|---------------|
| Confidence rating | Hispanic or Latino | 53 | 89 | 12.01 |
| | Not hispanic or Latino | 28 | 81 | 20.66 |
| | DNA | 2 | 88 | 13.58 |
| ML agreement | Hispanic or Latino | 53 | 87 | 13.04 |
| | Not hispanic or Latino | 28 | 90 | 6.66 |
| | DNA | 2 | 88 | 2.43 |

learning techniques). We first tested this approach using data that were previously collected (and coded live by human observers) in the lab and then applied the approach to data collected remotely.

Our results revealed that the overall agreement between the live observers and the ML model was high (~90% for 6- and 36-month-olds) suggesting that the combination of OpenFace and ML performs at a level similar to well-established methods for collecting looking time data. After demonstrating that OpenFace could be used to estimate gaze for infants and children in a lab setting, we expanded our dataset to include children tested remotely in the home. Because these data could not be coded live by observers in real time, we used attention getting stimuli to compare the ML model's predictions of gaze locations to known locations. Once again, the ML model's estimates of gaze locations had high agreement (~90% for 48-, 60-, and 72-month-olds) with that of the known locations on the screen, demonstrating that this method is suitable for estimating gaze direction for data collected remotely.

FUTURE DIRECTIONS AND LIMITATIONS

While we have demonstrated initial success in implementing this novel approach, it is important to note that both data collection and model development are ongoing. Further, we should acknowledge, that while results of our ML model are promising, we are just beginning to test its effectiveness with infants. Preliminary results look promising. Thus far, we have tested four infants online ($n = 2$ at 15 months, $n = 1$ at 13 months, and $n = 1$ at 4 months). We have processed their video data in OpenFace. Mean confidence ratings were

as follows: 15-month confidence ratings = 90.39%, 87.46%, 13-month confidence rating = 70.63%, 4-month confidence rating = 68.73%. While the 15-month data look strikingly similar to the average of our online data, the confidence ratings for the 13- and 4-month-old infants were slightly lower. Again, we should also note, there were no initial criteria in place for data quality of the videos processed in OpenFace. As such, we are confident that with further development, once factors such as looking away from the screen or the face being obscured are taken into account, confident ratings will increase. Future research should take this into account.

One current limitation of our approach is that the ML model requires training on a subset of the data. However, our ultimate goal is to establish a fully autonomous model that is robust enough to classify gaze without training on a subset of the data. As our data set continues to grow, so too will the generalizability of our ML model. In addition, we plan to train more complex models with a larger number of parameters once we incorporate additional data from multiple participants. This involves adding more layers and more neurons to our neural network. Specifically, we will use K-fold cross validation for training the model. This works by randomly dividing the training data set into groups (folds) and repeating training steps K times (where K = the number of iterations) while at each time we hold out one of the sections of data (folds). This allows us to test the accuracy of the model on multiple sections of the dataset, increasing our overall accuracy estimates.

Another limitation of our current approach is that it was developed specifically for coding data from a three-screen audio-visual protocol (the MAAP). Once we incorporate all the data from multiple participants for training, the ML model will be able to be used for various types of input (i.e., other looking time tasks), and can be scaled up to incorporate more complex gaze estimates (e.g., more than three locations). Our lab is currently adapting a ML model to be used with the Intersensory Processing Efficiency Protocol (IPEP; Bahrick et al., 2018b), an audiovisual task similar to the MAAP, but with six AOIs as opposed three.

Recently, the approach that has been outlined in this manuscript has been adapted by the Multisensory Data Network (a collection of 13 research labs across North America) who will administer the MAAP and IPEP remotely, process the looking time data using OpenFace and our ML model, and add to our growing dataset, helping to further inform and refine our model as well as develop preliminary norms for the development of skills assessed by the MAAP and IPEP. We plan on publishing our dataset (once it is complete) for scientists to use. As such, this paper provides an important first step in developing an open source toolkit capable of quantifying large-scale looking time data collected remotely for infants and children.

Finally, we acknowledge that the minimum technological requirements of a home computer with a web camera and high-speed internet connection could potentially limit our participant pool. Specifically, as Lourenco and Tasimi (2020) have recently pointed out, the technological requirements of many online studies may restrict access to many low-income and minority communities and thus, may impact the generalizability of the findings.

CONCLUSIONS

These preliminary results suggest that under the right circumstances, OpenFace can be used with infants (both with data collected in a lab setting and preliminary data collected remotely) and with young children (for tasks that were administered remotely) to derive gaze vectors for looking time. Further, when paired with our ML model, we can accurately and efficiently process looking time data and provide an output that is comparable in accuracy to traditional methods of looking time. As such, our approach provides developmental researchers with a viable option for collecting looking time data outside of the typical laboratory setting. Not only does this provide researchers with a cost-effective method for data collection, but it also frees them from the geographical confines of testing individuals within the typical university community, opening the door to a world-wide participant pool.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Florida International University Office of Research Integrity. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individuals and minor(s) legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

BE, JT, and LB developed the study concept. BE, JT, AS, and LB contributed to the study design. EE, VP, MM, and WG performed data collection and coding. BE, JT, and AS performed the data analysis and interpretation. BE drafted the manuscript, JT, AS, and LB provided critical revisions. All authors contributed to and approved the final version of the manuscript for submission.

FUNDING

This research was supported in part by the National Institute on Minority Health and Health Disparities of the National Institutes of Health Under Award Number NIMHD (U54MD012393), Florida International University Research Center in Minority Institutions, awarded to BE. This work was also supported by the NICHD: Grants R01-HD053776 and R01-HD094803 awarded to LB.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.731618/full#supplementary-material>

REFERENCES

- Akinyelu, A. A., and Blignaut, P. (2020). Convolutional neural network-based methods for eye gaze estimation: a survey. *IEEE Access* 8, 142581–142605. doi: 10.1109/ACCESS.2020.3013540
- Altvater-Mackensen, N., Mani, N., and Grossmann, T. (2016). Audiovisual speech perception in infancy: the influence of vowel identity and infants' productive abilities on sensitivity to (mis)matches between auditory and visual speech cues. *Dev. Psychol.* 52, 191–204. doi: 10.1037/a0039964
- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., and Evershed, J. K. (2020b). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behav. Res. Methods*. 53, 1407–1425. doi: 10.3758/s13428-020-01501-5
- Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. (2018). Gorilla in our MIDST: an online behavioral experiment builder. *BioRxiv* 2018, 388–407. doi: 10.1101/438242
- Anwyl-Irvine, A. L., Dalmaijer, E., Hodges, N., and Evershed, J. (2020a). Online timing accuracy and precision: a comparison of platforms, browsers, and participant's devices. *PsyArXiv*. 2020, 1–22. doi: 10.31234/osf.io/jfeca
- Aslin, R. N. (2012). Infant eyes: a window on cognitive development. *Infancy* 17, 126–140. doi: 10.1111/j.1532-7078.2011.0097.x
- Bahrack, L. E. (1987). Infants' intermodal perception of two levels of temporal structure in natural events. *Infant Behav. Dev.* 10, 387–416. doi: 10.1016/0163-6383(87)90039-7
- Bahrack, L. E., and Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Dev. Psychol.* 36, 190–201. doi: 10.1037/0012-1649.36.2.190
- Bahrack, L. E., Soska, K. C., and Todd, J. T. (2018b). Assessing individual differences in the speed and accuracy of intersensory processing in young children: the intersensory processing efficiency protocol. *Dev. Psychol.* 54, 2226–2239. doi: 10.1037/dev0000575
- Bahrack, L. E., Todd, J. T., and Soska, K. C. (2018a). The Multisensory Attention Assessment Protocol (MAAP): characterizing individual differences in multisensory attention skills in infants and children and relations with language and cognition. *Dev. Psychol.* 54, 2207–2225. doi: 10.1037/dev000594
- Bahrack, L. E., and Watson, J. S. (1985). Detection of intermodal proprioceptive-visual contingency as a potential basis of self-perception in infancy. *Dev. Psychol.* 21, 963–973. doi: 10.1037/0012-1649.21.6.963
- Baltrusaitis, T., Robinson, P., and Morency, L. P. (2016). "OpenFace: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016* (Lake Placid, NY). doi: 10.1109/WACV.2016.7477553
- Casey, B., and Richards, J. E. (1988). Sustained visual attention in young infants measured with an adapted version of the visual preference paradigm. *Child Dev.* 59, 1514–1521.
- Chouinard, B., Scott, K., and Cusack, R. (2019). Using automatic face analysis to score infant behaviour from video collected online. *Infant Behav. Dev.* 54, 1–12. doi: 10.1016/j.infbeh.2018.11.004
- Colombo, J., Mitchell, D. W., Coldren, J. T., and Freese, L. J. (1991). Individual differences in infant visual attention: are short lookers faster processors or feature processors? *Child Dev.* 62, 1247–1257.
- Emberson, L. L., Zinszer, B. D., Raizada, R. D. S., and Aslin, R. N. (2017). Decoding the infant mind: multivariate pattern analysis (MVPA) using fNIRS. *PLoS ONE* 12: e0172500. doi: 10.1371/journal.pone.0172500
- Fernald, A., Zangl, R., Portillo, A. L., and Marchman, V. A. (2008). "Looking while listening: using eye movements to monitor spoken language," in *Developmental Psycholinguistics: On-Line Methods in Children's Language Processing*, eds I. A. Sekerina, E. M. Fernández, and H. Clahsen (John Benjamins Publishing Company), 97–135.
- George, A., and Routray, A. (2016). Fast and accurate algorithm for eye localisation for gaze tracking in low-resolution images. *IET Comput. Vis.* 10, 660–669. doi: 10.1049/iet-cvi.2015.0316
- Goodwin, L. D., and Leech, N. L. (2006). Understanding correlation: factors that affect the size of r. *J. Exp. Educ.* 74, 249–266. doi: 10.3200/JEXE.74.3.249-266
- Hessels, R. S., and Hooge, I. T. C. (2019). Eye tracking in developmental cognitive neuroscience – The good, the bad and the ugly. *Dev. Cogn. Neurosci.* 40: 100710. doi: 10.1016/j.dcn.2019.100710
- Hirsh-Pasek, K., and Golinkoff, R. M. (1996). "The intermodal preferential looking paradigm: a window onto emerging language comprehension," in eds *Language, Speech, and Communication. Methods for Assessing Children's Syntax*, D. McDaniel, C. McKee, and H. S. Cairns (The MIT Press), 104–124.
- Hutton, S. B. (2019). "Eye tracking methodology," in *Eye Movement Research. Studies in Neuroscience, Psychology and Behavioral Economics*, eds C. Klein and U. Ettinger (Cham: Springer), 277–308.
- Ioffe, S., and Szegedy, C. (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift", in *International Conference on Machine Learning*, 448–456.
- Jaccard, J. J., and Becker, M. A. (2009). *Statistics for the Behavioral Sciences*, 5th ed. Boston, MA: Wadsworth Publishing.
- Jesse, A., and Johnson, E. K. (2016). Audiovisual alignment of co-speech gestures to speech supports word learning in 2-year-olds. *J. Exp. Child Psychol.* 145, 1–10. doi: 10.1016/j.jecp.2015.12.002
- Klaib, A. F., Alsrhein, N. O., Melhem, W. Y., Bashtawi, H. O., and Magableh, A. A. (2021). Eye tracking algorithms, techniques, tools, and applications with an emphasis on machine learning and Internet of Things technologies. *Expert Syst. Appl.* 166: 114037. doi: 10.1016/j.eswa.2020.114037
- Lewkowicz, D. J. (1988). Sensory dominance in infants: II. Ten-month-old infants' response to auditory-visual compounds. *Dev. Psychol.* 24, 172–182. doi: 10.1037//0012-1649.24.2.172
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv[Preprint]*. arXiv:1908.09635. doi: 10.1145/3457607
- Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," In *ICML*.
- Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430. doi: 10.1016/j.tics.2006.07.005
- Oakes, L. M. (2010). Infancy guidelines for publishing eye-tracking data. *Infancy* 15, 1–5. doi: 10.1111/j.1532-7078.2010.0030.x
- Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy* 17, 1–8. doi: 10.1111/j.1532-7078.2011.00101.x
- Park, S., Spurr, A., and Hilliges, O. (2018). Deep pictorial gaze estimation. *LNCS* 11217, 741–757. doi: 10.1007/978-3-030-01261-8_44
- Richards, J. E. (1987). Infant visual sustained attention and respiratory sinus arrhythmia. *Child Dev.* 58, 488–496.
- Ross-Sheehy, S., Reynolds, E., and Eschman, B. (2021). Unsupervised online assessment of visual working memory in 4-to 10-year-old children: array size influences capacity estimates and task performance. *Front. Psychol.* 12, 2410. doi: 10.3389/fpsyg.2021.692228
- Ross-Sheehy, S., Schneegans, S., and Spencer, J. P. (2015). The infant orienting with attention task: assessing the neural basis of spatial attention in infancy. *Infancy* 20, 467–506. doi: 10.1111/infa.12087
- Scott, K., and Schulz, L. (2017). Lookit (Part 1): a new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/opmi_a_00002
- Semmelmann, K., and Weigelt, S. (2017). Online eye tracking with consumer-grade webcams: potential and limits. *J. Vis.* 17, 892–892. doi: 10.1167/17.10.892

- Shaddy, D. J., and Colombo, J. (2004). Developmental changes in infant attention to dynamic and static stimuli. *Infancy* 5, 355–365. doi: 10.1207/s15327078in0503_6
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.
- Wass, S. V., Smith, T. J., and Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behav. Res. Methods* 45, 229–250. doi: 10.3758/s13428-012-0245-6
- Wood, E., and Bulling, A. (2014). “EyeTab: model-based gaze estimation on unmodified tablet computers,” in *Eye Tracking Research and Applications Symposium (ETRA)* (Safety Harbor, FL), 207–210. doi: 10.1145/2578153.2578185

Author Disclaimer: The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Eschman, Todd, Sarafraz, Edgar, Petrulla, McNew, Gomez and Bahrck. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Zoom, Zoom, Baby! Assessing Mother-Infant Interaction During the Still Face Paradigm and Infant Language Development *via* a Virtual Visit Procedure

Nancy L. McElwain^{1,2*}, Yannan Hu¹, Xiaomei Li¹, Meghan C. Fisher¹, Jenny C. Baldwin¹ and Jordan M. Bodway¹

¹ Department of Human Development and Family Studies, College of Agricultural, Consumer, and Environmental Sciences, University of Illinois at Urbana-Champaign, Urbana, IL, United States, ² Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, United States

OPEN ACCESS

Edited by:

Lisa Oakes,
University of California, Davis,
United States

Reviewed by:

Natalie Brito,
New York University, United States
Erica Neri,
University of Bologna, Italy

*Correspondence:

Nancy L. McElwain
mcelwn@illinois.edu

Specialty section:

This article was submitted to
Developmental Psychology,
a section of the journal
Frontiers in Psychology

Received: 01 July 2021

Accepted: 27 December 2021

Published: 16 February 2022

Citation:

McElwain NL, Hu Y, Li X, Fisher MC,
Baldwin JC and Bodway JM (2022)
Zoom, Zoom, Baby! Assessing
Mother-Infant Interaction During the
Still Face Paradigm and Infant
Language Development *via* a Virtual
Visit Procedure.
Front. Psychol. 12:734492.
doi: 10.3389/fpsyg.2021.734492

The COVID-19 pandemic has necessitated innovations in data collection protocols, including use of virtual or remote visits. Although developmental scientists used virtual visits prior to COVID-19, validation of virtual assessments of infant socioemotional and language development are lacking. We aimed to fill this gap by validating a virtual visit protocol that assesses mother and infant behavior during the Still Face Paradigm (SFP) and infant receptive and expressive communication using the Bayley-III Screening Test. Validation was accomplished through comparisons of data (i.e., proportions of missing data for a given task; observed infant and maternal behaviors) collected during in-person laboratory visits and virtual visits conducted *via* Zoom. Of the 119 mother-infant dyads who participated, 73 participated in lab visits only, 13 participated in virtual visits only, and 33 dyads participated in a combination of lab and virtual visits across four time points (3, 6, 9, and 12 months). Maternal perspectives of, and preferences for, virtual visits were also assessed. Proportions of missing data were higher during virtual visits, particularly for assessments of infant receptive communication. Nonetheless, comparisons of virtual and laboratory visits within a given time point (3, 6, or 9 months) indicated that mothers and infants showed similar proportions of facial expressions, vocalizations and directions of gaze during the SFP and infants showed similar and expected patterns of behavioral change across SFP episodes. Infants also demonstrated comparable expressive and receptive communicative abilities across virtual and laboratory assessments. Maternal reports of ease and preference for virtual visits varied by infant age, with mothers of 12-month-old infants reporting, on average, less ease of virtual visits and a preference for in-person visits. Results are discussed in terms of feasibility and validity of virtual visits for assessing infant socioemotional and language development, and broader advantages and disadvantages of virtual visits are also considered.

Keywords: infant stress, language development, mother-infant interaction, virtual visits, participant experience

INTRODUCTION

The COVID-19 pandemic and related restrictions have created challenges for developmental research that relies heavily on traditional in-person methods of data collection. Yet, in meeting those challenges, and building on psychological researchers' successful use of online testing platforms with adults, adolescents, and school-age children (e.g., Buhrmester et al., 2011; Germine et al., 2012; Griffiths et al., 2019), developmental scientists have explored and fine-tuned creative and potentially transformative solutions to conducting research with infants and young children. Although online methods were in use by developmental researchers prior to COVID-19 (e.g., Scott and Schulz, 2017; Tran et al., 2017), the need for such methods during the pandemic has spurred further development and proliferation (see Garrisi et al., 2020; Su and Ceci, 2021). The majority of online or virtual validation studies, to date, have focused on cognitive developmental assessments, whereas validated virtual assessments of infant socioemotional and language development have been sparse. We aimed to fill this gap.

As with almost all facets of life—research and otherwise—across the globe, our longitudinal investigation of infant development was halted in March 2020. We quickly pivoted to a virtual visit protocol using a video conferencing platform, which resulted in a unique opportunity to compare laboratory and virtual visits in assessing infant socioemotional and language functioning. Specifically, we assessed (a) infant and maternal behavior and infant response to stress during the Still Face Paradigm (SFP; Tronick et al., 1978) using a micro-behavioral coding approach, and (b) infant language development using the expressive and receptive communication subtests of the Bayley-III Screening Test (Bayley, 2006a). We also examined maternal perceptions of, and preferences for, virtual vs. laboratory visits.

Our study complements innovative efforts by cognitive developmentalists to collect data *via* online platforms. Asynchronous unmoderated platforms, such as LookIt (Scott and Schulz, 2017; Scott et al., 2017) and Amazon Mechanical Turk (Tran et al., 2017), have been used to conduct infant looking-time studies. Although asynchronous studies with infants and young children have demonstrated feasibility (e.g., Scott and Schulz, 2017; Tran et al., 2017; Rhodes et al., 2020), large portions of data (e.g., 40% in Tran et al., 2017) are typically excluded due to technical problems or procedural errors (e.g., infant position in the video). Online studies conducted synchronously (i.e., with a live experimenter present) have demonstrated feasibility, with minimal data loss, in assessing neurodevelopmental risk (Kelleher et al., 2020), looking time and learning (Smith-Flores et al., 2021) and cognition and memory (Sheskin and Keil, 2018) among infants and children. Importantly, both asynchronous and synchronous online studies of varied cognitive domains largely yield findings that replicate those from laboratory studies (e.g., Scott and Schulz, 2017; Sheskin and Keil, 2018; Rhodes et al., 2020; Smith-Flores et al., 2021).

Although this growing literature suggests the promise of online platforms for assessing cognitive development in the context of highly structured tasks, the validity on such methods for assessing dimensions of infant social and emotional

functioning, such as parent-infant interaction, infant response to stress, and infant expressive and receptive communication skills, remains unknown. Whereas certain advantages (e.g., greater flexibility, more diverse participant pool) and disadvantages (e.g., poor internet connectivity, decreased experimental control, increased potential for distractions, see Su and Ceci, 2021) regarding virtual visits will be common across studies of cognitive and socioemotional development, some issues are unique. On the one hand, virtual visits may be particularly conducive to capturing infant and maternal social and emotional behaviors that are more ecologically valid because assessments take place in the familiar home environment without experimenters physically present. On the other hand, and precisely because the home environment is highly familiar, the effectiveness of virtual visits in eliciting infant response to stress may be reduced. Additionally, intensive assessments of mother-infant interaction using microanalytic coding schemes typically require video recording procedures that involve multiple cameras and pan/zoom/tilt functionality. Although two recent studies indicate feasibility of administering mother-infant interaction tasks *via* a virtual visit protocol (Gustafsson et al., 2021; Shin et al., 2021), feasibility was assessed subjectively (e.g., research assistant ratings), and validity was not assessed. With these issues in mind, we examined objective indicators of feasibility and validity of synchronous virtual visit procedures designed to assess mother-infant interaction, infant stress regulation, and infant language development.

With respect to infant socioemotional functioning, the Still-Face Paradigm (SFP; Tronick et al., 1978) is an established procedure to assess mother-infant interaction and infant responses to stress. The SFP involves three episodes, each typically 2 mins in length: (a) a “play” episode, in which the mother and infant interact without toys, (b) a “still face” episode, in which the mother looks at her infant but maintains a neutral facial expression and ceases interaction (i.e., vocalizations, touch), and (c) a “reunion” episode, in which the mother resumes interaction with her infant. The still face episode, which violates infants' expectations for reciprocal interaction, typically elicits a distress response. Use of microanalytic coding procedures, in which infant and maternal behaviors are coded continuously, permits a window into maternal and infant behavioral coordination during the play and reunion episodes (e.g., Sravish et al., 2013; Pratt et al., 2015; Busuito and Moore, 2017). Further, a meta-analysis of 39 studies indicated robust and expected effects of the SFP on infant behavior coded in a microanalytic manner: infant gaze to mother and positive affect decreased and infant negative affect increased from the play to still face episode, whereas infant positive affect and gaze to mother increased from still face to reunion (Mesman et al., 2009).

Importantly, Mesman et al. (2009) reported that SFP effects on infant behavior and affect were robust to procedural differences (e.g., length of episodes, use of transition interval in which mother turned away from infants between episodes) across studies, although consideration of the setting (i.e., lab vs. home) in which the procedure was carried out was not considered in this meta-analytic review. Whereas most studies using the SFP have been conducted in a controlled laboratory environment,

Moore et al. (2001) conducted the SFP during home visits at 2, 4, and 6 months and reported expected changes in infant behavior (i.e., increases in negative affect and decreases in positive affect during the still face episode), thus supporting the use of the SFP in the home environment. Moreover, Gustafsson et al. (2021) reported that among 348 mother-infant dyads participating in a virtual visit procedure, including the SFP, 94–99% of videos passed data quality checks based on research assistant ratings. Although promising, objective evidence of feasibility (e.g., percentage of missing data) and validity of virtual SFP assessments is lacking.

In contrast to the lack of prior validation studies for remote assessments of mother-infant interaction during the SFP, several prior studies have reported feasibility and validity of assessing child language abilities using an online video conferencing format. Findings indicate that speech and language characteristics (e.g., mean length utterance, number of different words) among toddlers during play with a parent (Manning et al., 2020) and performance on a standardized language assessment among school-age children with language impairment (Sutherland et al., 2017) showed good feasibility, and reliability and/or validity of assessments did not differ significantly from data collected during face-to-face sessions. Ashworth et al. (2021), however, reported significantly higher verbal performance (assessed *via* the British Picture Vocabulary Scale, Third Edition [BPVS-3]) during online virtual visits vs. laboratory visits among school-aged children with Williams syndrome. Although these past studies indicate utility of conducting language assessments among toddlers and school-aged children *via* a virtual visit platform, we are unaware of prior work that has compared infant language assessments conducted *via* a synchronous virtual visit vs. an in-person laboratory format.

Complementing direct assessments of infant socioemotional and language functioning, assessing parents' perspectives about their virtual visit experiences is also needed. To date, the pros and cons of virtual visits for developmental research have been primarily discussed from the perspective of the researcher (see Su and Ceci, 2021). In this vein and in our own experience, advantages of virtual visits include greater re/scheduling flexibility, greater efficiency of cost and time, and better ability to recruit more (geographically) diverse samples, whereas disadvantages include diminished researcher control, greater dependence on parents to implement task procedures, and technical challenges that arise due to poor internet connectivity and/or shortcomings of participants' devices. Equally important, however, are parents' views about participation in developmental research using online platforms. Although there has been long-standing interest in clinical research (e.g., Yessis et al., 2012) and health care settings (e.g., Cleary and Edgman-Levitan, 1997) for assessing participants' or patients' perspectives of their experiences, such assessments are less common in developmental research (but see Kelleher et al., 2020; Maitre et al., 2021). Given the relative novelty of remote assessment methods in developmental research, and particularly the lack of such methods used to assess infant socioemotional and language development, mothers' perceptions of and preferences for visits of this type may provide a window into how and when

such visits may be best used, as well as input for refining visit procedures in ways that not only increase data quality but also optimize participants' experience.

In the current study, we addressed three main objectives. First, we assessed whether infant and maternal SFP behaviors differed between laboratory and virtual visits on several objective metrics, including the frequency of missing data, distributions of behavioral codes, and expected changes in infant behavior across episodes. Second, we assessed whether infant receptive and expressive language differed between laboratory and virtual visits on similar metrics, including frequency of missing data and infant subtest scores. Third, we assessed mothers' perceptions of the virtual visit format and their preferences for virtual visits vs. in-person laboratory visits. We also conducted supplementary analyses to examine whether (a) the behavioral variables assessed during virtual visits and (b) maternal perceptions and preferences for virtual visits varied as a function of the dyad's prior experience with virtual visits.

METHOD

Participants

One hundred and nineteen infants (57 girls; 48%) and their mothers participated in a short-term longitudinal study from 3 to 12 months of age, in which the overarching goal was to investigate mother-infant interaction dynamics and attachment formation as predictors of infant physiological regulation and brain development. Families were recruited from local pediatric clinics, community organizations, and online forums serving families from a wide range of socioeconomic and racial/ethnic backgrounds. Families were excluded from participating if their infant had any known cardiac abnormalities, was born preterm (<37 weeks gestation), had birth complications and/or admission to the NICU, or had an MRI contraindication. Mothers were 13% Asian, 7% Black or African American, 73% White non-Hispanic, 4% Hispanic and 3% another race or more than one race. Forty-percent of mothers had completed a bachelor's degree, and 39% had received an advanced degree. The average annual family income was between \$61,000 and \$70,000.

Sample sizes and descriptive statistics for infant age and sex as a function of visit type at each time point are reported in **Table 1**. As shown in **Table 1**, 73 dyads participated in lab visits only, 13 dyads participated in virtual visits only, and 33 dyads participated in a combination of lab and virtual visits (labeled "hybrid" visit schedule). Chi-square analyses comparing visit type (lab vs. virtual visit at a given time point) by infant sex revealed one significant difference: At the 3-month time point, a higher proportion of female infants participated in virtual visits compared with laboratory visits, $\chi^2(1) = 4.20$, $p = 0.04$. No other differences emerged for infant sex as a function of visit type. Because the one infant sex difference that emerged was based on small cell sizes (10 females vs. 3 males), and because no differences in the main study measures differed as a function of lab vs. virtual visits at 3 months (see Results), we did not consider infant sex further in our analyses. Additionally, *t*-tests for independent samples revealed no difference for infant age at each time point as a function of visit type, and one-way

TABLE 1 | Sample sizes and infant characteristics for mother-infant dyads participating in lab visits only, virtual visits only, and hybrid visits (lab and virtual).

| Visit schedule | 3 months | 6 months | 9 months | 12 months |
|--------------------------------------|-------------|-------------|-------------|--------------|
| Lab visits only (<i>n</i> = 73) | 49 | 67 | 62 | 58 |
| Infant sex (% female) | 59% | 52% | 52% | 50% |
| Infant mean age (<i>SD</i>) | 3.22 (0.29) | 6.22 (0.38) | 9.33 (0.43) | 12.68 (0.48) |
| Virtual visits only (<i>n</i> = 13) | 13 | 12 | 11 | 12 |
| Infant sex (% female) | 77% | 75% | 82% | 75% |
| Infant mean age (<i>SD</i>) | 3.34 (0.41) | 6.28 (0.28) | 9.26 (0.30) | 12.38 (0.23) |
| Hybrid visits (<i>n</i> = 33) | 33 (0) | 26 (7) | 13 (19) | 0 (32) |
| Infant sex (% female) | 27% | 27% | 25% | 28% |
| Infant mean age (<i>SD</i>) | 3.33 (0.35) | 6.22 (0.25) | 9.34 (0.35) | 12.89 (0.91) |
| Total visits (<i>N</i> = 119) | 95 | 112 | 105 | 102 |

For hybrid visits, the number of dyads participating in lab visits at a given time point is shown first, followed by the number of dyads participating in virtual visits shown in parentheses.

ANOVAs with visit schedule (lab only, virtual visit only, hybrid) as the between-subjects factor revealed no significant differences in maternal education and family income.

Overview of Study Procedures

Laboratory Visits

Prior to COVID-19, infants and mothers participated in a 60-mins laboratory visit at 3, 6, and 9 months and a 90-mins visit at 12 months; infant brain scans (via magnetic resonance imaging [MRI]) during natural sleep were also conducted at 3 and 12 months. The laboratory visits included assessments of behavior and physiology (using 3-lead ECG wireless monitors) during a baseline session, play session, challenging puzzle task (12 months only), SFP (3, 6, 9 months only; Tronick et al., 1978), Strange Situation Procedure (12 months only; Ainsworth et al., 1978), as well as administration of Bayley cognitive and language subtests. Behavioral data from the SFP and Bayley language subtests were examined in this report. For the SFP, the infant was seated in an age-appropriate seat (e.g., bouncy seat, high chair). Mothers were provided with both verbal and written instructions about the SFP episodes, and a gentle knock on the door to the laboratory playroom signaled when to transition to the next episode. Following the mother-infant interaction sessions, a trained research assistant administered the Bayley-III Screening Test. Two professional cameras were mounted in opposite corners of the playroom; cameras had pan/tilt/zoom capabilities and were controlled and viewed from an observational booth adjacent to the playroom. All tasks were recorded for later review or scoring. Parents also completed a series of online questionnaires at each time point *via* Qualtrics.

Virtual Visits

During a 40-min virtual visit, we conducted assessments paralleling our laboratory assessments of (a) infant baseline physiology, (b) mother-infant play, (c) SFP (3, 6, 9 months only), (d) challenging puzzle task (12 months only), and (e) Bayley language subtests. Prior to the virtual visit, mothers were emailed a Zoom link as well as a list of materials needed during the visit (e.g., bouncy seat or high chair for SFP) and were

reminded to charge the device (e.g., laptop, tablet, phone) they planned to use for the visit (see Garrisi et al., 2020; Smith-Flores et al., 2021, regarding recommendations for Zoom as platform for synchronous virtual visits). Zoom links were sent with the passcode function to protect participant privacy. Host and participant videos were switched to “on” and the waiting room feature was enabled. The visit coordinator recorded the session directly to the local computer (vs. Zoom cloud option) and data were subsequently uploaded to secure servers. The visit coordinator used the share screen function in Zoom to present slides that detailed instructions for each activity. After giving an overview of the visit activities and informing mothers of their right to request that the session or recording be stopped at any time (as was also done at the beginning of the laboratory visits), the visit coordinator started video recording and pinned the participant’s video. During all activities, the visit coordinator turned off her video camera and microphone (except during the Baseline video- unmuted) and asked mothers to minimize their Zoom window so that infants would not be distracted by the screen. The share screen function was also used to share (a) the sea animals video for the baseline physiological assessment and (b) picture items for the Bayley receptive communication subtest. The visit coordinator worked with the mother to obtain the optimal video angle for each activity (e.g., capturing faces of both the mother and infant).

Participants were video recorded in a variety of rooms including living rooms, dining rooms, parent and infant bedrooms, and kitchens. Nonetheless, virtual visits maintained consistency with the lab visits in that an infant bouncy seat or high chair was used for the SFP, with mothers seated on the floor or in a chair accordingly. To provide an optimal camera angle for behavioral coding of the SFP, mothers and infants sat facing each other and slightly angled themselves toward the camera so that their faces were visible. To proactively minimize distractions, mothers were asked to silence their phones, turn off any electronic toys used during the play sessions, and limit the presence of pets or other family members as much as possible. Following the virtual visit, parents completed online questionnaires, which included a brief survey about maternal perceptions of the virtual visit.

Measures

Behavioral Coding of the Still Face Paradigm (SFP)

The SFP was micro-coded for infant and maternal facial expressions, vocalizations, and directions of gaze using Datavyu 1.4.1, which allows for onset/offset coding of behaviors in real time and segmenting continuous codes by frame (33ms per frame). The current codes were adapted from existing coding systems on mother-infant interactions (Tronick et al., 1980; Braungart-Rieker et al., 1998; Moore et al., 2001; Moore and Calkins, 2004). For each code, categories were mutually exclusive and exhaustive. Data were coded as missing when the behavior of interest was not visible or audible, or when there was an interruption. With the following exception, the coding system and procedures were identical across laboratory and virtual visits. For the virtual visits only, we included “gaze at screen” as another category in the gaze code (see below) for both mothers and

infants to capture degree to which the device used for the virtual visit was a distraction.

Infant Facial Expression

Codes were cry, frown, unalert (e.g., sleepy, yawn), alert neutral (e.g., wary, sober, bright, coo face), mild positive (e.g., simple or subtle smile), strong positive (e.g., broad smile, appearance of laughter), and other (e.g., sneeze, cough). *Infant vocalization* codes were cry, fuss, positive neutral (e.g., babble, coos), laughter, yawn, other (e.g., sneeze, cough, burp), and none. *Infant direction of gaze* codes were gazing at mother's face, gazing at mother's actions (e.g., following mother's moving fingers), gazing at other objects (e.g., gazing at mother's static legs or high chair), gazing away, eyes closed, and gazing at screen (virtual visits only).

Mother Facial Expression

Codes were angry, distressed, flat, interested, simple smile, broad smile, and other (e.g., yawn, sneeze, cough, sniff). *Mother vocalization* codes were infant-direct speech (i.e., baby talk with much change of modulation, bigger jump in pitch, or elongated vowels), adult speech (i.e., speaking normally as if she is talking to an adult with "regular" rhythm and intonation), playful noise (e.g., "raspberries" kissing sounds, animal sounds, "bounce-bounce-bounce"), rhythmic sounds or singing, laughter, demanding speech, whisper, other (e.g., yawn, sneeze, cough, grunt, sigh), and none. *Mother direction of gaze* codes were gazing toward infant's face, gazing at infant's body or interaction-related objects, gazing away, and gazing at screen (virtual visits only).

Two separate teams coded infant and maternal behaviors, and coders who participated in coding the laboratory visits also coded the virtual visits. Coders went through intensive training for one to three months and gained high reliability ($kappa \geq 0.80$) on each code before proceeding with coding individual tapes. Within each coding team, different coders assessed behavior during the SFP across the three visits (3, 6, 9 months) whenever possible. Interobserver reliability, computed for 19–25% of tapes that spanned across the entire coding process, were consistently satisfactory ($kappa > 0.60$; see McHugh, 2012) across all codes and time points (see **Table 2**). Of note, reliability was calculated using by-frame output where each unit represent 33 ms. A time buffer of 10 frames ($\sim 1/3$ sec) was used to adjust for slight differences in coders' reaction times, and frames coded as missing by one or both coders were excluded from reliability calculations.

Infant Language Development

To assess infant receptive and expressive communication, we used the Bayley Scales of Infant and Toddler Development, Third Edition, Screening Test (Bayley, 2006a). The Bayley-III Screening Test is made up of items from the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley, 2006b) and is a well-established and validated screening instrument designed to assess cognitive, language, and motor functioning of infants and young children (Bayley, 2006a). This screening instrument was normed on a US representative sample of 1,675 children aged 1–42 months and demonstrates excellent test-retest reliability,

TABLE 2 | Interobserver reliability statistics (Cohen's kappa) for infant and maternal behaviors in the Still Face Paradigm separately by visit type.

| Behavioral codes | 3 months | | 6 months | | 9 months | |
|---------------------------|----------|---------|----------|---------|----------|---------|
| | Lab | Virtual | Lab | Virtual | Lab | Virtual |
| Infant behaviors | | | | | | |
| <i>N</i> (%) double-coded | 17 (22%) | 3 (23%) | 18 (20%) | 4 (22%) | 15 (21%) | 5 (19%) |
| Facial expression | 0.74 | 0.76 | 0.69 | 0.72 | 0.72 | 0.77 |
| Vocalization | 0.79 | 0.67 | 0.83 | 0.76 | 0.83 | 0.70 |
| Direction of gaze | 0.76 | 0.70 | 0.82 | 0.74 | 0.78 | 0.73 |
| Maternal behaviors | | | | | | |
| <i>N</i> (%) double-coded | 18 (23%) | 3 (23%) | 19 (21%) | 4 (22%) | 18 (25%) | 5 (19%) |
| Facial expression | 0.77 | 0.85 | 0.77 | 0.81 | 0.75 | 0.86 |
| Vocalization | 0.85 | 0.87 | 0.86 | 0.89 | 0.82 | 0.86 |
| Direction of gaze | 0.71 | 0.86 | 0.78 | 0.74 | 0.72 | 0.86 |

adequate internal consistency among subscale items, and good construct validity (Bayley, 2006a,b).

At both the virtual and laboratory visits, subtests assessing receptive communication (24 items) and expressive communication (24 items) were administered by a trained research assistant. The cognitive subtest was administered during the laboratory visits as well, but it was not deemed feasible for the virtual visits due to the required testing manipulatives, number of items, and complexity of instructing mothers to effectively administer the items. For infants in the age range participating in this study (3–12 months), items assessing receptive communication captured auditory acuity (e.g., responding to voices, discriminating sounds, localizing sounds), vocabulary development and comprehension (e.g., identifying objects or pictures that are referenced), and social referencing. Items assessing expressive communication captured preverbal communication (e.g., babbling, gesturing, joint referencing, turn taking) and vocabulary development (e.g., imitating words, naming pictures).

Following the Bayley (2006a) procedures, the infant's age determined the items to be administered, and administration ended when the infant failed four consecutive items. During the laboratory administration, the mother was present in the room. Administration for the language subtests took ~ 10 –15 mins, although we note that the majority of the expressive communication items did not require formal administration and could instead be scored through "incidental observation" (i.e., observing the infant during any part of the laboratory or virtual visit for expressive behaviors that satisfied the scoring criteria for a given item).

Several adaptations in administration procedures were made for the virtual visits. Prior to the visit, the visit coordinator informed the mother about materials needed for testing; these materials were typical items that a family with an infant would have at home (e.g., blocks, ball, spoon). To administer the Bayley-III Screening language subtests virtually, the visit coordinator (who had received extensive training in administration and scoring procedures) would observe the infant's verbal and

communicative behavior throughout the virtual visit to score items relevant to expressive communication and would guide the mother through the standard administration of items that could not be scored through incidental observation. In doing so, the visit coordinator provided verbal and written (through PowerPoint slides displayed *via* shared screen) instructions. For items in which a stimulus book was needed, we created comparable testing stimuli using open-source images available online, and the visit coordinator would show the pictures in PowerPoint slides *via* the share screen function in Zoom. The mother would verbally prompt the infant (e.g., *Show me the bird.*) The mother was instructed to indicate to the visit coordinator whether the infant was pointing or clearly looking at the item she mentioned. To minimize distractions, the visit coordinator also instructed mothers to minimize the Zoom screen during items where it was not needed.

For both laboratory and virtual visits, the researcher scored the receptive communication items as they were administered and reviewed the video recording following a given visit for any items for which questions arose. Scoring of expressive communication items varied slightly during laboratory vs. virtual visits. During the laboratory visits, the researcher administering the Bayley observed the infant *via* a video monitor during the different sessions (e.g., baseline, play, SFP), made detailed notes and scored items on the Bayley-III scoring form while carrying out these “live” observations of the infant; video recordings were reviewed following the visit if questions arose about scoring specific items. During the virtual visits, because one researcher managed all roles (e.g., providing mothers with task instructions, managing the video recording, administering Bayley items), the expressive items were scored almost exclusively by reviewing the video recording following the visit; paralleling scoring procedures from the laboratory visits, detailed notes were also entered on the scoring form. Further, a doctoral-level researcher with seven years of Bayley-III experience reviewed all Bayley procedures for all laboratory and virtual visits and corrected scores as needed to ensure accuracy of administration and scoring.

Maternal Perceptions of Virtual Visits

Following research visits at each time point, mothers completed a series of questionnaires online. Shortly after the onset of the virtual visits, we added items about maternal perceptions of the virtual visits to the questionnaire protocol, which mothers also completed at each time point in which they participated in a virtual visit. Mothers rated the following two items on a five-point scale: (a) Was the virtual visit easy for you to manage? (1 = *very easy* to 5 = *very challenging*), and (b) How well do you think the virtual visits vs. in-person visits capture you and your baby’s typical interactions? (1 = *much more typical during the virtual visit* to 5 = *much more typical during the in-person visit*; N/A option available for families who had not experienced an in-person visit). Mothers were also asked whether they would choose a virtual or in-person visit if they had the choice (forced choice: *virtual*, *in person*, *no preference*). If a preference for virtual or in-person visits was indicated, the mother was asked to describe her reason/s for the given preference.

Data Analytic Plan

Below we outline the analyses conducted to address the three main aims of this report. In addition, supplementary analyses are presented after the main analyses and assessed whether infant and maternal data collected during virtual visits at 6, 9, and 12 months varied as function of infant–mother dyads’ prior experience with virtual visits within the context of this study.

Infant and Maternal Behavior (SFP)

Our first research objective was to assess the feasibility (i.e., rates of missing data) and validity (i.e., proportional breakdown of categories within a given code, as well as change in key infant behaviors) of the virtual visits vs. lab visits in capturing infant and maternal behavior during the SFP. We assessed two types of missing data for mother–infant behavioral data from the SFP. First, at each time point, we compared proportions of data from laboratory vs. virtual visits that were completely missing (i.e., the observational assessment was attempted or fully conducted but could not be coded) using *z* tests. Second, for SFP sessions that were deemed codeable, we considered the proportion of missing data for a given code when parts of the recording were not codeable due to the camera angle (e.g., infant or mother’s face out of camera view) or interruptions (e.g., a third person entered the room; a noise distracted baby or mother). Because the proportion scores showed high levels of skewness and kurtosis, we used the Mann–Whitney *U*-test for independent samples to compare the distributions of missingness for lab and virtual visits; this nonparametric test, which does not assume normal distribution of the dependent variable, was more appropriate than an independent samples *t*-test.

Next, to test whether infant and maternal behaviors varied by visit type, we first computed proportion scores for maternal and infant behaviors as the number of frames for a given behavior (e.g., gazing at mother’s face) divided by the total number of frames coded for that particular behavioral category (e.g., infant direction of gaze), excluding frames that were coded as missing. We then tested whether proportion scores were different across lab and virtual visits using Mann–Whitney *U*-test *s* for independent samples. Given the multiple comparisons made (3 episodes \times 3 time points), we applied a Bonferroni correction of $p < 0.006$ (0.05 divided by 9) for each infant and maternal code (facial expression, vocalization, direction of gaze).

Lastly, to assess whether expected patterns of change in infant negative affect, positive affect, and gaze varied as a function of visit type (and in accordance with prior work, see Mesman et al., 2009), we computed the following composite scores within episode and time point: *infant negative affect* (i.e., mean of infant negative facial affect [cry + frown] and vocalization [cry + fuss] proportion scores), *infant positive affect* (i.e., sum of infant mild positive and strong positive facial affect proportion scores), and *infant gaze toward mother* (i.e., sum of infant gaze at mother’s face and gaze at mother’s actions). To be comparable to lab visits, we excluded the “gaze to screen” code assessed during virtual visits when computing proportion scores for infant gaze. Repeated measures ANOVA were conducted separately for each dependent variable with SFP episode as the repeated/within-subjects factor and visit type as the between-subjects factor. Because infants’

participation in lab vs. virtual visits changed across time, it was not possible to include time point as a second repeated measure and, thus, separate models were tested at 3, 6, and 9 months.

Infant Receptive and Expressive Communication (Bayley-III Screening Test)

Our second research objective was to assess the feasibility (i.e., rates of missing data) and validity (i.e., infant performance) of the virtual visits vs. lab visits in capturing infant receptive and expressive communication. For each of the language subtests and at each time point, we compared proportions of data from lab vs. virtual visits that were completely missing (i.e., the Bayley subtest was attempted or fully administered but could not be scored due to missing items) using z tests. Second, to compare scores from lab and virtual visits, we conducted t -tests for independent samples by subtest and time point. For both sets of tests, we used a Bonferroni correction of $p < 0.0125$ (0.05 divided by 4 time points) to adjust for multiple comparisons across time.

Maternal Perceptions and Preferences

Our third objective was to assess maternal perceptions of the virtual visit format and preferences for virtual vs. in-person, lab visits. Across all time points, 104 survey responses were obtained from 44 mothers. Using all maternal report data available, we conducted single sample t -tests (for maternal “ease” and “typical” ratings) at each time point to determine whether mean ratings significantly differed from the midpoint of the scale (i.e., value of 3 = “neutral” or “about the same”). With respect to maternal visit preferences (prefer virtual visit, prefer lab visit, no preference), we conducted a one-sample proportion test at each time point to assess whether the proportion of mothers who reported preference for virtual and lab visits, respectively, differed significantly from 0.33 (the proportion expected by chance).

RESULTS

Infant and Maternal Behavior During the Still Face Paradigm Missing Data

We assessed two types of missing data. The proportions of cases completely missing SFP data at each time point are shown in **Table 3**. At each time point, a chi-square analysis compared rates of missing data by visit type (lab vs. virtual). Although all comparisons were nonsignificant, reasons for missingness varied by visit type. As shown in **Table 3**, infant distress was a frequent reason for missingness during lab visits, in particular. With respect to SFP data that were coded, we also considered proportion of missing data for a given code at a given time point due to sections of the session that could not be coded (e.g., poor camera angle, mother blocked infant from view). As shown in **Table 4**, data were missing at higher proportions during virtual vs. lab visits, with 10 of 18 tests indicating a significant difference by visit type. Differences in proportions of missingness emerged for infant facial expression at all three time points and for mother facial expressions at two of the three time points. Facial expression missingness was also greatest in absolute terms at all time points and for both infants and mothers.

TABLE 3 | Dyads with data missing for the Still Face Paradigm as a function of visit type.

| Reason for missingness | Total <i>N</i> | 3 months | 6 months | 9 months |
|------------------------|----------------|----------|----------|----------|
| Lab visits | 19 (7.6%) | 10 (12%) | 4 (4%) | 5 (7%) |
| Equipment failure | 4 | 2 | 1 | 1 |
| Infant distress | 12 | 7 | 3 | 4 |
| Time constraints | 1 | 1 | 0 | 0 |
| Virtual visits | 4 (6.5%) | 0 (0%) | 1 (5%) | 3 (10%) |
| Experimenter error | 1 | 0 | 1 | 0 |
| Infant distress | 1 | 0 | 0 | 1 |
| Screen distraction | 2 | 0 | 0 | 2 |

TABLE 4 | Proportion of missing data during the Still Face Paradigm by behavioral code, time point and visit type.

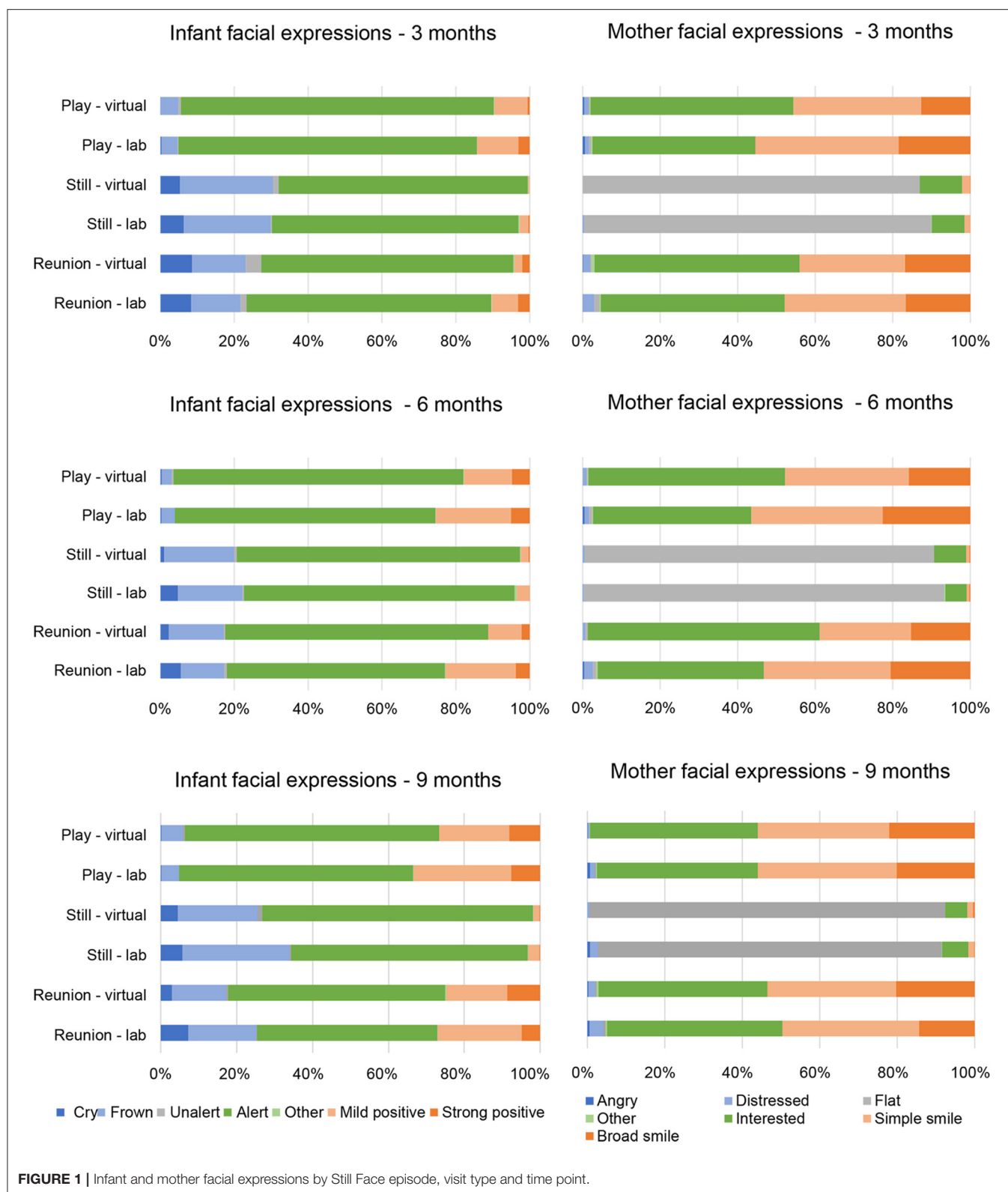
| Time point | Lab visits | Virtual visits | | |
|--------------------------|------------------------|------------------------|--------------------|----------------|
| Behavioral code | <i>M</i> (<i>SD</i>) | <i>M</i> (<i>SD</i>) | <i>z</i> statistic | <i>p</i> value |
| 3 months | | | | |
| Infant facial expression | 0.004 (0.018) | 0.103 (0.163) | 4.66 | <0.001 |
| Infant vocalization | 0.001 (0.004) | 0.006 (0.016) | 1.13 | 0.259 |
| Infant gaze | 0.004 (0.017) | 0.066 (0.113) | 3.92 | <0.001 |
| Mother facial expression | 0.030 (0.051) | 0.123 (0.117) | 3.36 | <0.001 |
| Mother vocalization | 0.001 (0.004) | 0.013 (0.022) | 3.19 | 0.001 |
| Mother gaze | 0.012 (0.029) | 0.029 (0.040) | 1.27 | 0.205 |
| 6 months | | | | |
| Infant facial expression | 0.017 (0.040) | 0.093 (0.117) | 4.44 | <0.001 |
| Infant vocalization | 0.009 (0.077) | 0.001 (0.003) | −0.13 | 0.895 |
| Infant gaze | 0.009 (0.025) | 0.025 (0.063) | 2.40 | 0.016 |
| Mother facial expression | 0.059 (0.078) | 0.131 (0.206) | 1.70 | 0.089 |
| Mother vocalization | 0.010 (0.077) | 0.005 (0.009) | 1.49 | 0.135 |
| Mother gaze | 0.029 (0.065) | 0.019 (0.031) | −0.54 | 0.593 |
| 9 months | | | | |
| Infant facial expression | 0.017 (0.038) | 0.078 (0.064) | 5.14 | < .001 |
| Infant vocalization | 0.002 (0.005) | 0.003 (0.008) | −0.80 | 0.423 |
| Infant gaze | 0.009 (0.013) | 0.022 (0.030) | 1.57 | 0.118 |
| Mother facial expression | 0.061 (0.066) | 0.128 (0.091) | 4.35 | <0.001 |
| Mother vocalization | 0.001 (0.002) | 0.006 (0.010) | 2.74 | 0.006 |
| Mother gaze | 0.031 (0.048) | 0.069 (0.111) | 2.98 | 0.003 |

*Mann-Whitney U-tests were performed to test differences between lab visits and virtual visits in proportion scores for a given code at a given time point. Standardized test statistics and *p* values are shown in the table.*

Distributions of Behavioral Codes

Within each episode of the SFP and each time point, we plotted the proportion of maternal and infant behaviors for facial expression (see **Figure 1**), vocalization (see **Figure 2**) and gaze (see **Figure 3**) codes, respectively.

As shown in **Figure 1**, “alert” was the most frequent infant facial expression across time points and SFP episodes. Mothers showed high frequencies of “interested,” “simple smile,” and “broad smile” during the play and reunion SFP episodes and high frequencies of “flat” during the still episode at each time point. Of the 63 Mann-Whitney U -tests conducted for infant



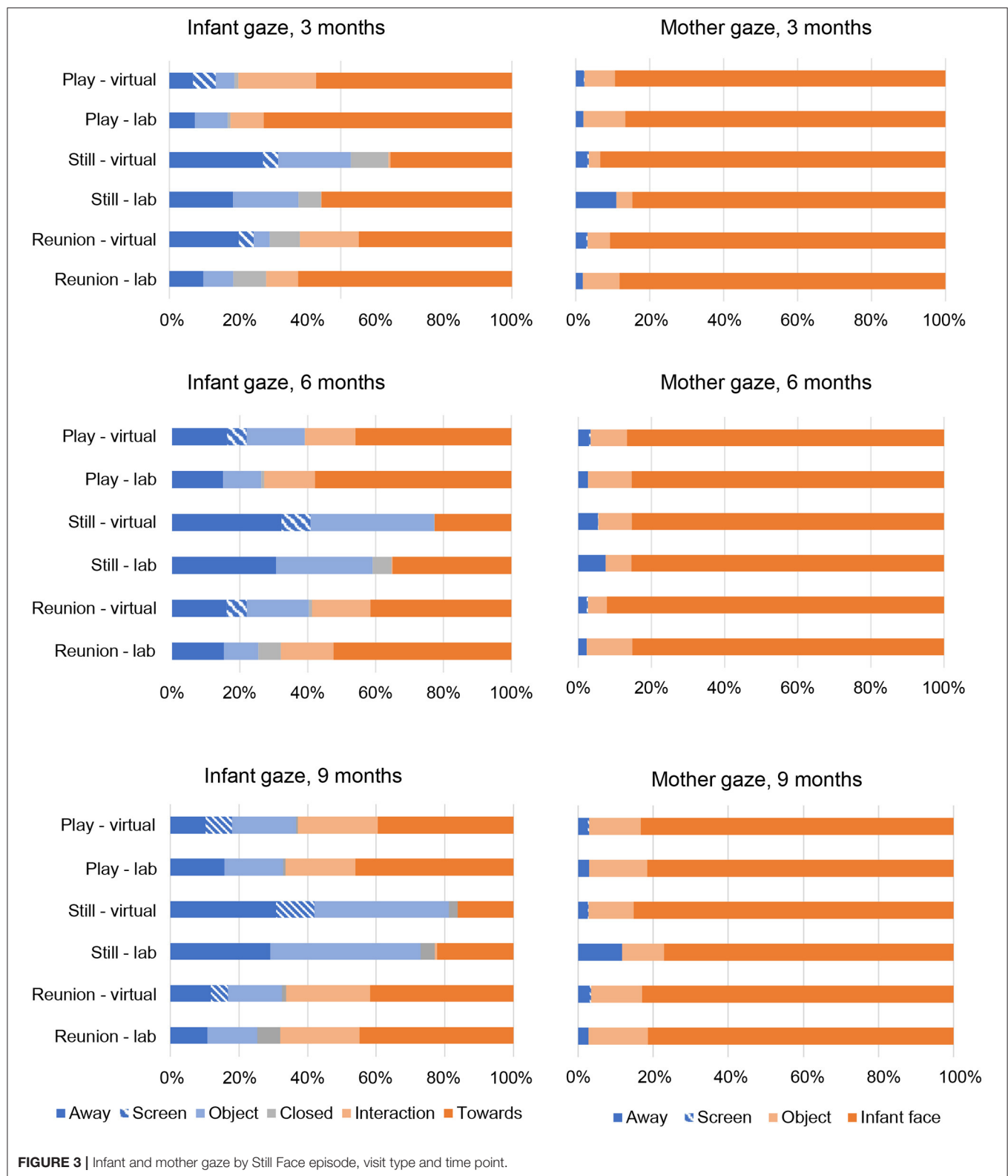
facial expressions (7 codes \times 3 episodes \times 3 time points), one comparison was significant at $p < 0.006$ (i.e., correction for multiple comparisons). At 6 months, infant showed more unalert

facial expressions in the play episode ($z = -3.89$, $p < 0.001$) during virtual vs. lab visits. Of the 63 tests conducted for mother facial expressions (7 codes \times 3 episodes \times 3 time points), one



was significant at $p < 0.006$. At 6 months, mothers showed more interested facial expressions ($z = -2.82, p = 0.005$) in the reunion episode during virtual vs. lab visits.

As shown in **Figure 2**, “none” was the most frequent infant vocalization code across time points and SFP episodes, whereas mothers showed relatively high frequencies of “infant-directed,”



“playful,” and “rhythmic” vocalizations during the play and reunion episodes at each time point. Of the 54 Mann-Whitney *U*-tests conducted for infant vocalizations (6 codes \times 3 episodes

\times 3 time points), one was significant at $p < 0.006$. Infants engaged in more crying during the still episode at 9 months ($z = -3.04$, $p = 0.002$) in lab vs. virtual visits. Of the 81 tests conducted for

mother vocalizations (9 codes \times 3 episodes \times 3 time points), two were significant at $p < 0.006$. Mothers were more likely to be silent (i.e., vocalizations coded as “none”) during the still episode at 6 months in virtual vs. lab visits ($z = -3.285$, $p = 0.001$) and more likely to exhibit “other” vocalizations (e.g., yawn, sneeze, cough) during the reunion episode at 9 months in lab vs. virtual visits ($z = -2.92$, $p = 0.003$).

As shown in **Figure 3**, and at each time point, infants’ gaze toward the mother’s face or actions was predominant during both the play and reunion episodes, whereas infant gaze was more evenly divided among gazing away, gazing at object, and gazing at mother’s face during the still episode. At each time point and across episodes, mothers showed high frequencies of gazing at infant’s face. Note that “gaze at screen” was only coded during the virtual visits to capture infant or maternal distraction with the device used for the virtual visit. Of the 45 Mann-Whitney U -tests conducted for infant gaze (5 codes [excluding “gaze at screen”] \times 3 episodes \times 3 time points), and the 27 tests conducted for maternal gaze (3 codes [excluding “gaze at screen”] \times 3 episodes \times 3 time points), none were significant at $p < 0.006$.

Change in Infant Behavior Across SFP Episodes

Next, we report mean proportion scores and 95% confidence intervals for infant negative affect (see **Figure 4**), infant positive affect (see **Figure 5**), and infant gaze toward mother (see **Figure 6**) as a function of SFP episode, visit type and time point. As shown in **Table 5**, the main effect of episode was significant at $p < 0.001$ for each infant composite at each time point, and the main effect of episode did not vary as a function of visit type (i.e., Episode \times visit type interaction was nonsignificant in all cases).

Post-hoc analyses of the episode main effect revealed expected changes in infant behavior, such that (a) infant negative affect was significantly lower during the play episode vs. still and

reunion episodes, with a difference between the still and reunion episodes (more negative affect in the still episode) emerging only at 9 months, (b) infant positive affect was significantly lower during the still episode vs. the play and reunion episodes, with no difference between these latter episodes, and (c) infant gaze toward mother was significantly lower during the still episode vs. play and reunion episodes, with a difference between play and reunion (less gaze toward mother during reunion) present only at 3 months. The main effect of visit type was significant in one instance. Across all episodes of the SFP, infants exhibited more positive affect during lab visits compared with virtual visits at the 6-month time point. All other main effects of visit type were nonsignificant.

Infant Receptive and Expressive Communication

Missing Data

Within each time point, comparisons of proportions of infants missing data on receptive or expressive language subtests at lab vs. virtual visits indicated significantly higher proportions of missingness on receptive language subtests at virtual vs. lab visits at 3, 6, and 9 months (see **Table 6**). Proportions of missing data on (a) receptive language at 12 months and (b) expressive language at any time point did not differ significantly by visit type. Infant distress/fatigue was the most common reasons for missing scores on the Bayley-III language subtests during the lab visits, whereas difficulty administering and/or scoring items (specific to the receptive language subtest) was the most common reason for missing data during the virtual visits. Additional although much less common reasons for missing language data for both lab and virtual visits included equipment failure, parental time constraints, and experimenter error.

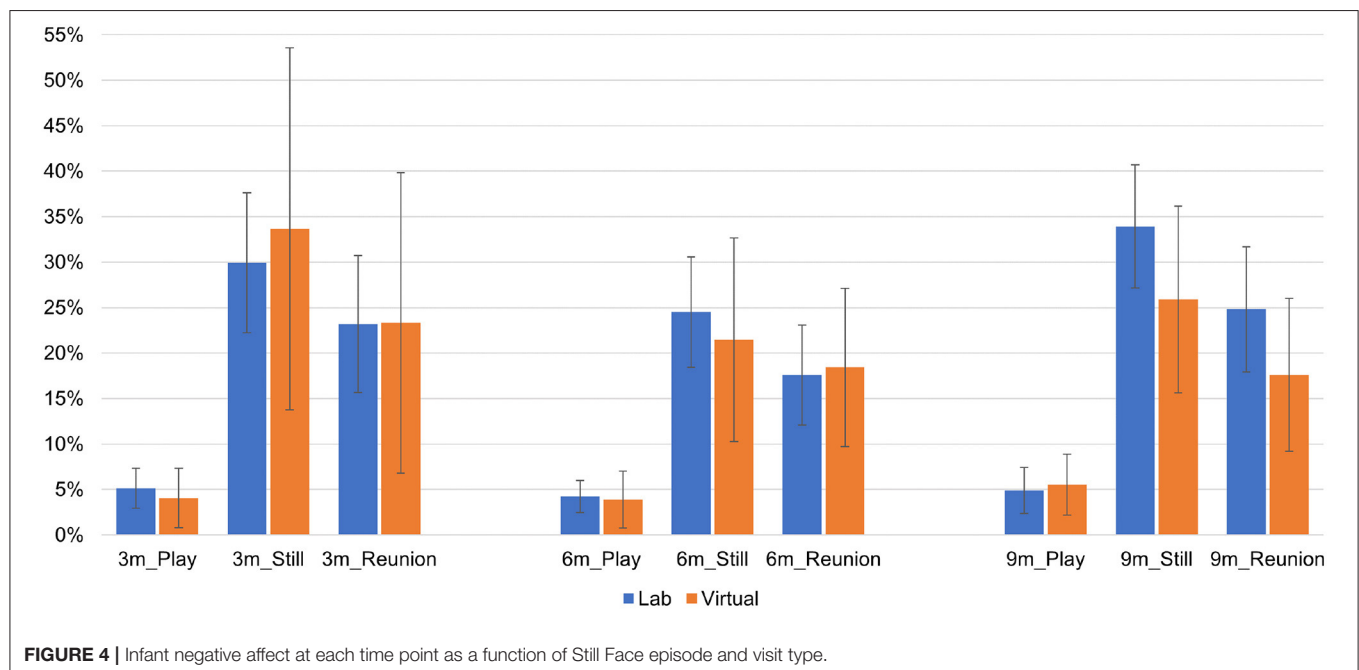


FIGURE 4 | Infant negative affect at each time point as a function of Still Face episode and visit type.

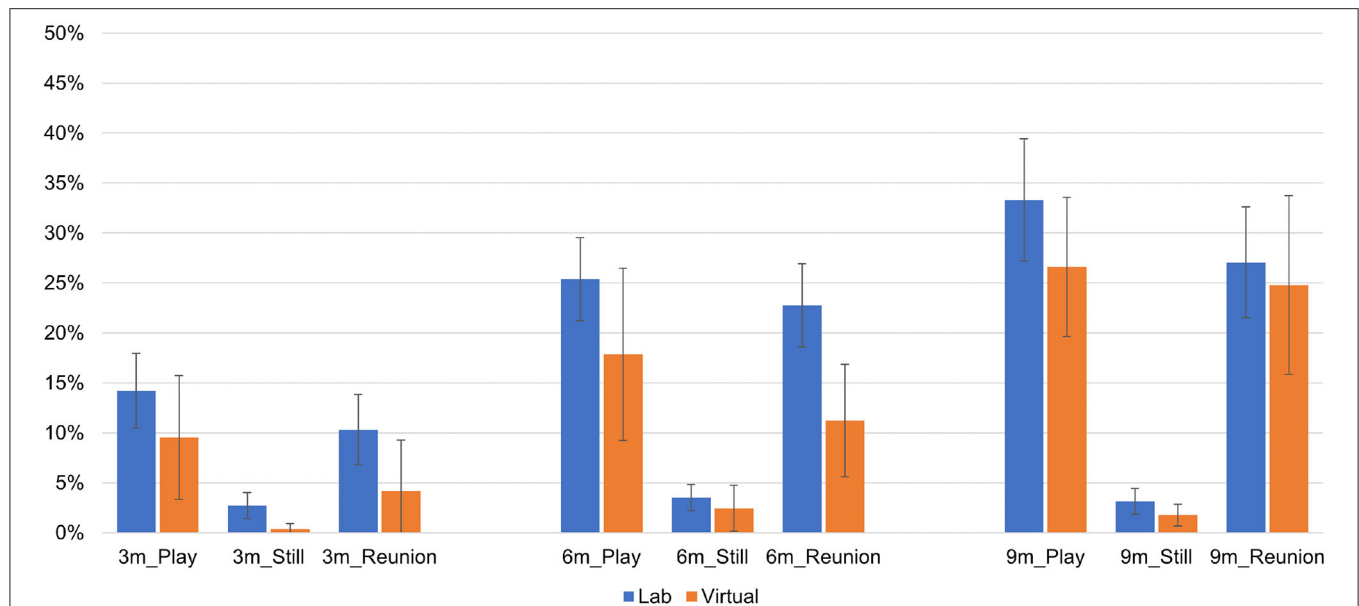


FIGURE 5 | Infant positive affect at each time point as a function of Still Face episode and visit type.

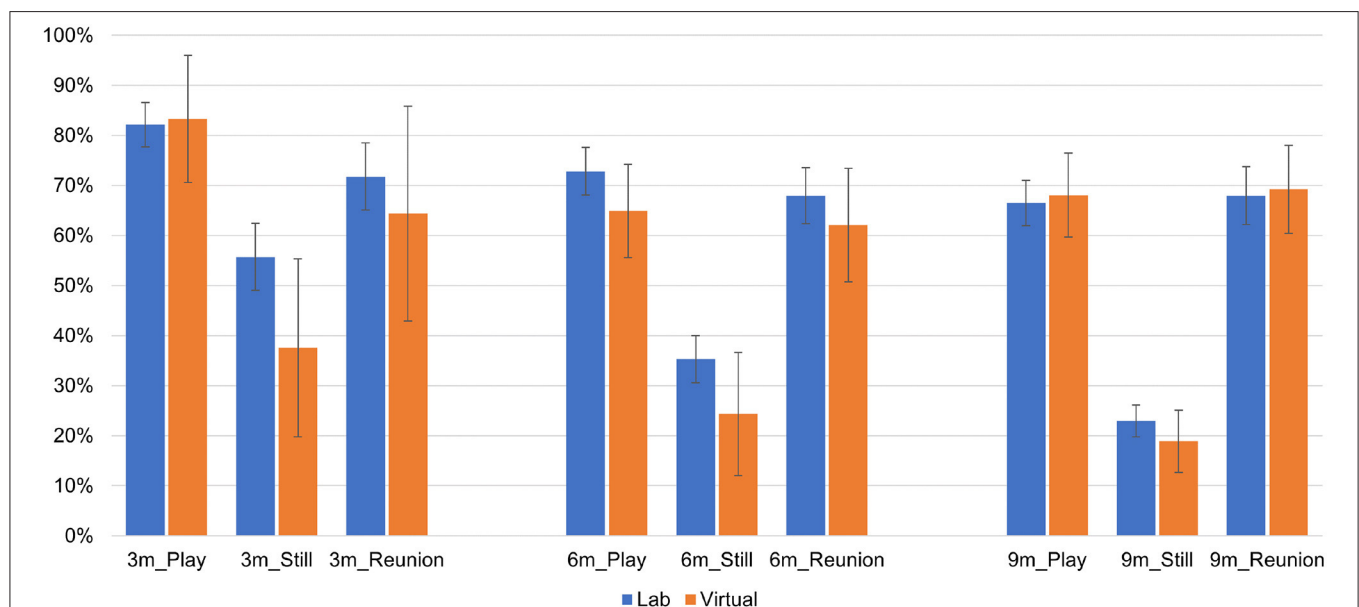


FIGURE 6 | Infant gaze to mother at each time point as a function of Still Face episode and visit type.

Receptive and Expressive Communication Scores

Mean and standard deviations for infant receptive and expressive communication scores are shown by time point and visit type in **Table 6**. Across all time points and for each subtest, *t*-tests for independent samples revealed no significant differences at $p < 0.0125$ (Bonferroni correction) in infant language scores as a function of virtual vs. lab visits (see **Table 6**).

Maternal Perceptions and Preferences

Mothers' virtual visit ratings and visit preferences as a function of time point and visit type are shown in **Figure 7**. As a preliminary step, we assessed whether demographic characteristics (infant sex, maternal education, family income) were related to maternal virtual visit perceptions or preferences. We used maternal reports following the mother's first virtual

TABLE 5 | Tests of infant behavioral change at each time point as a function of SFP episode and visit type.

| Dependent variables | Effects | MS | df | F | p | Episode mean differences | | |
|---------------------|----------------------|-------|----|--------|--------|--------------------------|---------------|--------------|
| | | | | | | Still—Play | Still—Reunion | Reunion—Play |
| Negative affect | | | | | | | | |
| 3 months | Episode | 0.851 | 2 | 21.16 | <0.001 | 0.27*** | 0.09 | 0.19** |
| | Visit type | 0.003 | 1 | 0.02 | 0.884 | | | |
| | Episode × visit type | 0.007 | 2 | 0.17 | 0.845 | | | |
| 6 months | Episode | 0.576 | 2 | 18.70 | <0.001 | 0.19*** | 0.05 | 0.14*** |
| | Visit type | 0.003 | 1 | 0.038 | 0.847 | | | |
| | Episode × visit type | 0.006 | 2 | 0.19 | 0.824 | | | |
| 9 months | Episode | 1.224 | 2 | 37.25 | <0.001 | 0.25*** | 0.09** | 0.16*** |
| | Visit type | 0.139 | 1 | 1.48 | 0.227 | | | |
| | Episode × visit type | 0.045 | 2 | 1.37 | 0.258 | | | |
| Positive affect | | | | | | | | |
| 3 months | Episode | 0.118 | 2 | 10.63 | <0.001 | −0.10*** | −0.06* | −0.05 |
| | Visit type | 0.064 | 1 | 2.66 | 0.107 | | | |
| | Episode × visit type | 0.004 | 2 | 0.37 | 0.691 | | | |
| 6 months | Episode | 0.564 | 2 | 32.60 | <0.001 | −0.19*** | −0.14*** | −0.05 |
| | Visit type | 0.202 | 1 | 4.916 | 0.029 | | | |
| | Episode × visit type | 0.042 | 2 | 2.41 | 0.093 | | | |
| 9 months | Episode | 1.719 | 2 | 81.97 | <0.001 | −0.28*** | −0.24*** | −0.04 |
| | Visit type | 0.070 | 1 | 0.99 | 0.321 | | | |
| | Episode × visit type | 0.016 | 2 | 0.76 | 0.469 | | | |
| Gaze to mother | | | | | | | | |
| 3 months | Episode | 1.45 | 2 | 34.45 | <0.001 | −0.36*** | −0.21*** | −0.15** |
| | Visit type | 0.219 | 1 | 1.79 | 0.185 | | | |
| | Episode × visit type | 0.103 | 2 | 2.44 | 0.090 | | | |
| 6 months | Episode | 2.772 | 2 | 84.96 | <0.001 | −0.39*** | −0.35*** | −0.04 |
| | Visit type | 0.307 | 1 | 2.98 | 0.087 | | | |
| | Episode × visit type | 0.010 | 2 | 0.30 | 0.742 | | | |
| 9 months | Episode | 5.750 | 2 | 272.52 | <0.001 | −0.46*** | −0.48*** | 0.01 |
| | Visit type | 0.001 | 1 | 0.014 | 0.907 | | | |
| | Episode × visit type | 0.020 | 2 | 0.929 | 0.397 | | | |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

visit (regardless of time point, $n = 44$), and all associations were nonsignificant.

Next, to assess whether maternal perceptions of virtual visits were significantly different than a “neutral” rating, single-sample t -tests were conducted within each time point and indicated a significant difference for maternal ratings of visit ease at each time point (see **Figure 7A**). Mothers, on average, rated the virtual visits on the “easy” end of the scale compared with the scale mid-point (3 = “neutral”): $t_{(12)} = -5.52$, $p < 0.001$ (3 months), $t_{(17)} = -5.83$, $p < 0.001$ (6 months), $t_{(28)} = -8.05$, $p < 0.001$ (9 months), and $t_{(43)} = -3.86$, $p < 0.001$ (12 months). For maternal ratings of how well the lab vs. virtual visits captured typical interactions between the mother and her infant (see **Figure 7B**), maternal mean ratings were not significantly different from the scale mid-point (3 = “about the same”).

One-sample proportion tests of maternal preferences (prefer virtual visit, prefer lab visit, no preference, see **Figure 7C**) indicated a significant difference in maternal preferences at 12

months only, with 72% of mothers indicating preference for in-person visits, whereas 11% and 16% indicated preference for virtual visits or no preference, respectively, $z = 5.40$, $p < 0.001$. If a visit preference was indicated, we also asked mother to provide a brief explanation for her preference. Preferences for virtual visits included (a) *convenience* for mothers and families to schedule a visit without disruption to family routines ($n = 10$), (b) *safety* of families and their decreased potential exposure to COVID-19 ($n = 6$), and (c) *familiarity*, with a small group of mothers reporting their infant’s behavior is more natural in the home environment vs. a new setting with strangers ($n = 5$). Preferences for in-person lab visits included (a) *decreased distractions* to mothers and infants ($n = 25$), (b) *greater confidence* in conducting visit procedures with the help of research staff, especially in handling technology ($n = 11$), and (c) *desire for face-to-face interaction* vs. interacting through a screen ($n = 8$). Mothers of older infants (12 months) were much more likely to report their preference for in-person lab visits due to

TABLE 6 | Infant receptive and expressive communication: missing data and infant performance as a function of visit type and time point.

| Language assessment | Lab visits | Virtual visits | | |
|---------------------|------------------------|------------------------|----------------|----------------|
| Missing data | <i>n</i> (%) | <i>n</i> (%) | <i>z</i> test | <i>p</i> value |
| Receptive language | | | | |
| 3 months | 6 (7%) | 4 (31%) | 2.56 | 0.010 |
| 6 months | 6 (7%) | 6 (32%) | 3.20 | 0.001 |
| 9 months | 5 (7%) | 8 (27%) | 2.81 | 0.005 |
| 12 months | 15 (26%) | 17 (41%) | 1.55 | 0.122 |
| Expressive language | | | | |
| 3 months | 2 (2%) | 0 (0%) | −0.57 | 0.569 |
| 6 months | 7 (8%) | 1 (5%) | −0.36 | 0.719 |
| 9 months | 3 (4%) | 1 (3%) | −0.16 | 0.872 |
| 12 months | 7 (12%) | 2 (5%) | −1.26 | 0.208 |
| Infant performance | <i>M</i> (<i>SD</i>) | <i>M</i> (<i>SD</i>) | <i>t</i> -test | <i>p</i> value |
| Receptive language | | | | |
| 3 months | 3.13 (1.15) | 2.78 (1.30) | −0.86 | 0.391 |
| 6 months | 5.93 (1.45) | 5.23 (0.93) | −1.69 | 0.094 |
| 9 months | 7.23 (1.25) | 7.09 (1.44) | −0.43 | 0.666 |
| 12 months | 9.00 (1.86) | 9.92 (2.77) | 1.64 | 0.107 |
| Expressive language | | | | |
| 3 months | 4.16 (1.08) | 4.46 (0.78) | 0.95 | 0.343 |
| 6 months | 5.24 (1.35) | 5.44 (0.92) | 0.63 | 0.533 |
| 9 months | 7.21 (1.81) | 7.86 (1.30) | 2.03 | 0.047 |
| 12 months | 10.69 (2.08) | 10.63 (2.08) | −0.14 | 0.890 |

decreased distractions and greater assistance by research staff in carrying out study procedures.

Given the difference that emerged in maternal preferences at 12 months, we conducted two sets of follow-up analyses. First, using mothers' reports completed after their first virtual visit only, we conducted one-way ANOVAs with time point as the between-subjects factor with maternal "ease" and "typical" ratings, respectively, as the dependent variable. These analyses allowed direct tests of whether maternal perceptions differed as a function of time point and also controlled for mothers' prior experience with virtual visits. For maternal ratings of the degree to which the virtual visit was easy vs. difficult, the main effect of time point was marginally significant, $F(3) = 2.37$, $p = 0.085$, although a planned contrast revealed that mothers reported less ease of virtual visits at 12 months ($M = 2.43$, $SD = 1.28$) compared with maternal reports at the other three time points combined (3 months: $M = 1.69$, $SD = 0.85$; 6 months: $M = 1.33$, $SD = 0.52$; 9 months: $M = 1.64$, $SD = 0.92$), $t_{(40)} = 2.65$, $p = 0.011$. No time-point effects emerged for maternal reports of how typical their interaction with their infant was during virtual vs. lab visits.

Second, we assessed whether maternal preferences for lab visits at the 12-month time point may have been due to COVID timing (e.g., the 12-month reports were collected later in the pandemic). To explore this possibility, we assessed whether there was a significant difference in COVID timing (computed as number of days between the date of the maternal report

and March 13, 2020, marking the beginning of COVID-related shutdown) as a function of visit time point, and this test was nonsignificant, $F(3) = 1.65$, $p = 0.182$. At each time point, we also conducted a one-way ANOVA to examine whether COVID timing differed as a function of maternal visit preferences, and all tests were nonsignificant.

Supplementary Analyses

Among the 19 dyads participating in a virtual visit at 6 months, 12 dyads had participated in a virtual visit at 3 months. Of the 30 dyads participating in a virtual visit at 9 months, 18 had participated in one or more prior virtual visits (11 dyads at 3 and 6 months; 7 dyads at 6 months). Of the 44 dyads participating in a virtual visit at 12 months, 29 had participated in one or more prior virtual visits (11 dyads at 3, 6, and 9 months; 7 dyads at 6 and 9 months; 12 dyads at 9 months). Given that participants differed in the degree to which they had previously experienced virtual visits, we conducted supplemental analyses that paralleled the main analyses reported above to assess whether key study variables (i.e., proportions of missing data, infant and maternal behaviors during the SFP, infant receptive and expressive language) differed as a function of prior virtual visit exposure (present vs. absent; 6-, 9-, and 12-month time points) or number of prior virtual visits (9- and 12-month time points). Correcting for multiple comparisons, no significant differences emerged as a function of prior virtual visit experience.

Maternal perceptions and preferences related to virtual visits were also examined as a function of prior exposure to virtual visits. We conducted a series of independent *t*-tests at each time point (6, 9 and 12 months, respectively) to assess whether maternal "ease" and "typical" ratings differed as a function of the mother's prior experience with virtual visits. At the 9- and 12-month time points, Kendall's tau-b (τ_b) correlations were tested to assess whether the number of prior virtual visits was associated with maternal ratings. All tests were nonsignificant. Chi-square tests were conducted to assess differences in visit preference by prior virtual visit experience. At 9 months, mothers indicated greater preference for lab visits when they had not had prior exposure to virtual visits (total $n = 11$, 73% lab, 0% virtual, 27% no preference), whereas preferences were more evenly distributed among mothers with prior experience (total $n = 18$, 17% lab, 39% virtual, 44% no preference), $\chi^2(2) = 10.47$, $p = 0.005$. Maternal visit preferences at 6 and 12 months did not significantly differ as a function of prior experience.

DISCUSSION

The COVID-19 pandemic has required researchers to adapt to a unique set of circumstances, and such adaptations are paving the way for innovative research methods that will likely continue to be used and developed. The current project adds to growing evidence for the feasibility and validity of virtual visits. Complementing existing online paradigms that have been used to assess infants' cognitive development (Scott et al., 2017; Tran et al., 2017; Smith-Flores et al., 2021), our study focused on the use of established assessments to capture infant socioemotional and language development during the first year of life. Further,

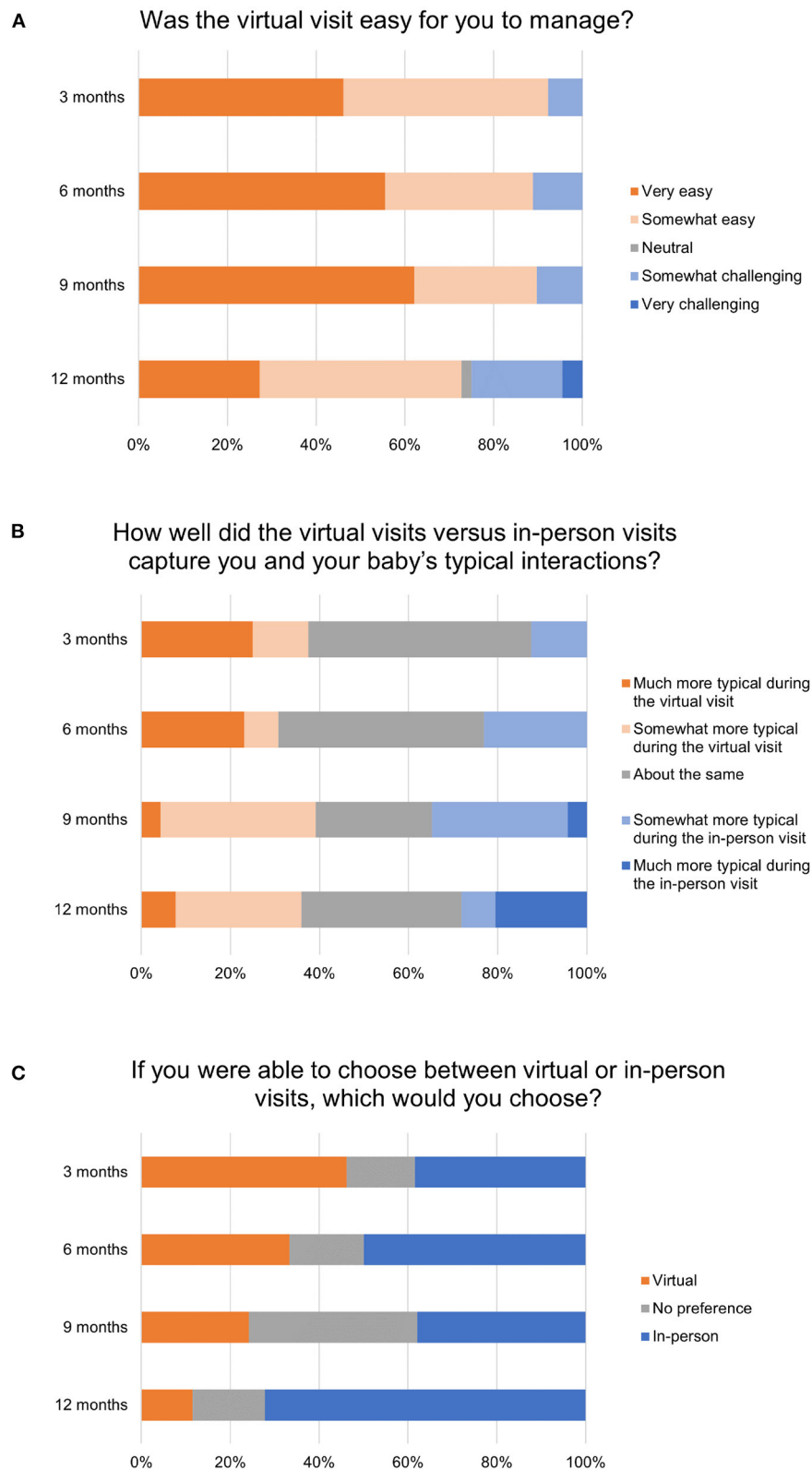


FIGURE 7 | Distributions of mothers' **(A)** ratings of ease of virtual visits, **(B)** ratings of how well the virtual visits captured typical interactions between the mother and her infant, and **(C)** preferences for virtual versus in-person visits. Exact wording of questionnaire items are shown above. Maternal responses to each item are displayed by time point.

by conducting assessments at multiple time points, we were able to explore whether virtual visits were more or less feasible or valid at different periods during the first year. Lastly, we gained insight into maternal perspectives on virtual visits and how such perspectives differed by infant age.

Our first objective was to assess the feasibility and validity of using the SFP in a virtual visit context. The SFP has demonstrated validity for assessing infant stress regulation in controlled laboratory settings (see Mesman et al., 2009), although Moore et al. (2001) provide evidence supporting use of the procedure in home environments. Our findings indicate that not only is it feasible to conduct the SFP using a virtual visit procedure (also see Gustafsson et al., 2021), but that applying a microanalytic coding scheme to video recordings from virtual visits yields data comparable to those collected in a controlled laboratory setting with professional camera capabilities. With respect to feasibility, the percentage of SFP protocols that were completely missing did not differ across lab and virtual visits, although reasons for missingness varied across visit type, with lab visits deemed not codeable largely due to technical issues or infant distress/fatigue (particularly among younger infants), whereas virtual visits were largely missing due to screen distraction among older infants. We note that we quickly adjusted our protocol (i.e., asking mothers to minimize their Zoom screen during the interactive tasks) when it became clear during our initial virtual visits that older infants were distracted by the computer screen. The convenience and scheduling flexibility of virtual visits also enabled us to minimize data loss due to infant distress or fatigue because we could more easily reschedule virtual visits when the mother indicated that another time would be better. These findings are consistent with prior online studies with a live experimenter that show minimal data loss (e.g., Sheskin and Keil, 2018), especially when task and recording/technical procedures have been thoroughly pilot tested and key privacy safeguards (e.g., waiting room, meeting password) are in place (e.g., see Garrisi et al., 2020, for a comprehensive set of recommendations).

Among cases that were deemed codeable, there were consistently higher levels of missing data during the virtual visits, as would be expected. In the laboratory playroom, two professional-grade cameras mounted in opposite corners of the room with zoom, pan, and tilt functions enabled high-quality recordings of mother and infant in split screen format. Nonetheless, despite the much more limited recording capabilities available during the virtual visits (i.e., only one camera angle was available for recording, and there was no ability to pan or tilt the camera to follow the mother and/or infant at moments when they moved out of the camera frame), missing data were relatively minimal and were highest for facial expressions (7–13% missingness, compared with <1–6% missingness for lab visits). Our visit coordinator worked with mothers to obtain the best recording possible, and the relatively low levels of missingness—both in terms of completely missing and partially missing—are acceptable, especially in comparison with levels of missing data observed for asynchronous sessions (e.g., Tran et al., 2017).

With respect to validity, we compared the proportions of behaviors for each infant and maternal code (i.e., facial

expressions, vocalizations, and directions of gaze) separately for SFP episodes and time points. Analyses indicated minimal differences in mean proportions. Of the 162 Mann-Whitney *U*-tests conducted for infant behaviors and the 171 tests conducted for maternal behaviors, 4 total were significant at $p < 0.006$ (Bonferroni correction). Further, using a microanalytic coding scheme in which behaviors were coded continuously on a frame-by-frame basis, our assessment of interobserver reliability on 19–25% of visits showed acceptable kappa statistics (>0.65) for all codes at all time points, regardless of visit type. Lastly, at each time point (3, 6, 9 months), infant negative affect, positive affect, and gaze to mother showed the expected patterns of change across play, still and reunion episodes of the SFP as reported in prior work (see Mesman et al., 2009). Importantly, the patterns of infant behavior change were largely consistent across time points, and in no instance did change patterns differ as a function of visit type. In prior studies, maternal and infant behavioral assessments during the SFP have typically been conducted in controlled laboratory settings (e.g., Sravish et al., 2013, but see Moore et al., 2001; Pratt et al., 2015; Busuito and Moore, 2017), raising questions about whether a virtual visit procedure carried out in the familiar home environment (and without researchers physically present) would elicit expected changes in infant affect and behavior. In addressing this concern, our study provides some of the first evidence for the feasibility and validity of assessing mother-infant interaction and infant behavioral regulation during the SFP using a virtual visit format.

Our second objective was to assess infant expressive and receptive communication using the Bayley-III Screening Test. With respect to infant performance, expressive and receptive communication scores at a given time point (3, 6, 9 and 12 months) did not differ as a function of visit type. These results are consistent with prior studies indicating no difference in language performance assessed *via* virtual vs. face-to-face visits among toddlers (Manning et al., 2020) and school-age children (Sutherland et al., 2017). Nonetheless, despite similar performance across the virtual and laboratory visits on the language subtests, there were significantly higher amounts of missing data on the receptive language subtest at the 3-, 6- and 9-month virtual vs. lab visits, which is cause for concern. Although rates of missing data did not significantly differ by visit type at 12 months, there were relatively high rates of missing data for both lab and virtual visits at this time point. The rates of missing data among the expressive subtest scores, in contrast, were relatively low at all time points for both virtual and lab visits.

Unlike the expressive language subtest, which relies mainly on researchers' "incidental observations" of the infant during the course of the virtual or lab visit, the receptive language subtest involves observing the infant's response to administered probes or items. We used stringent scoring criteria, in which subtest scores were considered to be completely missing if one or more items could not be adequately administered and/or scored due to infant compliance, distractions, and/or administrator error. Although we provided mothers with detailed instructions for administering the receptive language items, missing data during virtual visits at 3, 6, and 9 months was mainly due to faulty administration and related difficulty in scoring the infant

response. At 12 months, missing data at both virtual and lab visits were predominantly due to infant fatigue, noncompliance, and/or distractions. Taken together, our findings for the Bayley-III Screening communication subtests indicate high levels of feasibility and validity for assessing *expressive language* via a virtual visit procedure in which the infant and mother engage in a variety of interactive tasks that permit observing a range of infant expressive communication skills. Our confidence in virtual assessments of infants' expressive language specifically is corroborated by Manning et al.'s (2020) report on the validity of a virtual assessment of toddlers' language skills obtained during parent-toddler play session using indicators such as observed mean length utterance and number of different words spoken. Given the relatively large proportions of missing data for *receptive language*, however, we have less confidence in the feasibility of assessing this aspect of infant language development *via* a virtual visit procedure.

Our final objective was to assess maternal perspectives of, and preferences for, virtual visits using a brief survey created for the purposes of this study. At each time point, mothers were significantly more likely to rate virtual visits as "easy" compared with "neutral." Nonetheless, mothers of 12-month-olds were significantly more likely to rate virtual visits as less easy compared with maternal ratings at other infant ages. Mothers of 12-month-olds also showed a preference for in-person visits. Interestingly, although supplementary analyses indicated that prior experience with virtual visits in the context of the current study was not related to infant or maternal behavioral measures (infant or mother SFP behaviors; infant Bayley language scores), mothers of 9-month-olds reported a greater preference for in-person visits when they had not had prior virtual visit experience, whereas mothers of 12-month-olds indicated preference for in-person visits regardless of prior virtual visit experience. A substantial proportion of mothers of older infants (9- and 12-month-olds) also reported that their interactions with their infants were more natural during lab visits compared with virtual visits, although all comparisons on this item were nonsignificant.

Reasons for preferring in-person visits among mothers of older infants centered on distractions that surrounded virtual visits as well as mothers' desire for support by research staff. Given that 9- and 12-month-olds are "on the go" and more attuned to the wider environment, including the device used during the virtual visit, this pattern of maternal responses is perhaps not surprising. Yet, in-person lab visits may pose other challenges for older infants. Whereas it was possible for younger infants to take brief naps or feeding breaks during lab visits, such breaks were less feasible with older infants. As such, older infants may become more fatigued and fussier for some of the same reasons that mothers found virtual visits challenging. Because we did not collect mothers' ratings of relative ease or challenge of their experiences following in-person lab visits, we were unable to directly compare mothers' perceptions of virtual vs. lab visits. We were also unable to compare missingness, reliability, and validity of observational data from the 12-month visits because only data from the SFP sessions have been coded to date. In this light, we cannot make strong recommendations for the use of virtual visits at 12 months. In contrast, data at 3, 6, and 9 months suggest

that virtual visits are an acceptable, useful option for capturing infant socioemotional functioning observed during the SFP and infant expressive communication skills assessed *via* the Bayley-III Screening Test.

We note several limitations of the current study. First and foremost, we did not initially set out to assess the feasibility and validity of a virtual visit procedure. Instead, the study objectives in this report emerged as a result of necessary COVID-related restrictions that required us to pivot to a virtual visit protocol. Given the *ad-hoc* nature of the virtual visits, our sample sizes across visit types and time points were unbalanced, although we did conduct virtual visits at all time points to enable assessment of virtual visit feasibility and validity across a range of infant ages. Second, although few differences emerged between virtual and lab visits on our key study measures, COVID-19 posed a clear design confound. This confound was most concerning with respect to mothers of older infants showing a preference for lab vs. virtual visits. Follow-up analyses indicated that such preferences did not covary with when maternal reports were made relative to the COVID shutdown in March 2020. Nonetheless, more direct assessments of families' COVID-related stressors and experiences (e.g., disruptions to child care and work routines, illness, social isolation) are important to consider in relation to mothers' virtual visit perceptions and preferences. Likewise, certain advantages (e.g., scheduling flexibility) and disadvantages (e.g., distractions in the home) of virtual visits may be heightened due to the pandemic and become less salient in a post-pandemic environment.

In addition to limitations specific to COVID-19, the use of virtual visits to assess infant socioemotional and communicative competence requires consideration of broader advantages and disadvantages. Among older infants who are more mobile (crawling, walking), the lack of cameras that could follow the infant and often the lack of contained space in the home increased challenges of conducting assessments of mother-infant interactions with older infants. Further, distractions or interruptions from family members or pets during virtual visits can be a hindrance while carrying out standardized assessments, although we aimed to proactively limit distractions. During visits, mothers were also asked to minimize their Zoom windows during interactive tasks to minimize screen distractions. As already noted above, an important advantage of virtual visits was greater convenience and flexibility in scheduling and/or rescheduling visits as needed due to infant sleep schedule or mood. For instance, we were able to minimize such missing data during our virtual visits by offering mothers the opportunity to schedule a "catch-up" visit if their infant became fussy/tired. Although we provided this option at both lab and virtual visits, mothers were much more likely to schedule a catch-up virtual visit ($n = 32$) vs. a catch-up lab visit ($n = 15$). Not only does such flexibility in resuming virtual visits at a later time when infants are more alert decrease the likelihood of some types of missing data, it may also increase researchers' ability to more accurately capture infants' levels of competence.

We highlight two additional advantages of virtual visits that may provide further motivation for using such procedures to assess infant socioemotional development beyond the COVID-19

pandemic. First, virtual visits may provide a method to more effectively recruit fathers. Although we required mothers' participation for the current study to be consistent with our laboratory procedures, including fathers in virtual visit procedures will likely provide more flexibility to families and thereby increase participation rates. Research on the dynamics of parent-infant interaction and infant development has focused almost exclusively on mothers (e.g., see Davis and Logsdon, 2011), even though fathers have increasingly taken on caregiving roles over the past several decades and make unique contributions to children's socioemotional and cognitive outcomes (e.g., Cabrera et al., 2018; Ruiz et al., 2019). Research indicates, however, that on average, fathers who agree vs. decline to participate in high-commitment, high-stress research procedures (e.g., multiple videotaped procedures) differ on a host of factors, including education, race/ethnicity, infant characteristics, and family functioning (Costigan and Cox, 2001), suggesting the need for research approaches that are less burdensome to families and to fathers, in particular. Changes to family routines and stress on the whole family system brought about by the pandemic (Prime et al., 2020) further underscore the importance of capturing the larger family context and infants' experiences with mothers as well as fathers when present in the home. Virtual visits provide a novel, cost-effective, and family-friendly way to involve fathers more directly in research on infant development and family dynamics.

Second, and in a related vein, virtual visits may provide developmental researchers opportunities to recruit samples with greater racial, ethnic, socioeconomic and/or geographic diversity. To do so, however, researchers will need to be mindful of the "digital divide" faced by participants from rural communities and/or lower SES backgrounds and the obstacles they face in terms of access to reliable computers and internet connectivity (see Lourenco and Tasimi, 2020; van Dijk, 2020). In addition, families characterized by lower socioeconomic status may have less easy access to physical space with few distractions. As such, infant or dyad performance on virtual visit tasks could be impeded due to higher rates of technical issues and/or distractions, and further validation of virtual visit procedures among geographically and socioeconomically diverse samples is warranted.

In sum, the COVID-19 pandemic has accelerated the use of virtual visits by developmental researchers, and past work demonstrating validity of these techniques has predominantly focused on assessments of infant cognition. Our findings indicate that data obtained from assessments of infant socioemotional and language functioning using a synchronous virtual visit procedure

are comparable to those obtained during in-person lab visits. Although the use and validation of these new procedures during a global pandemic present inherent limitations, infant assessments conducted *via* Zoom and other remote platforms are likely to be used well beyond the current pandemic. Developmental researchers should continue to assess their feasibility and validity.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Office for the Protection of Research Subjects, University of Illinois at Urbana-Champaign. Written informed consent to participate in this study was provided by the participants' parents.

AUTHOR CONTRIBUTIONS

NM contributed to the conception and design of the study. NM, YH, XL, MF, and JCB contributed to the design of the virtual visit procedures. YH, XL, and MF contributed to oversight of data coding and scoring. NM and YH contributed to the statistical analyses and presentation of results in figures. XL, MF, JCB, and JMB contributed to writing sections of the manuscript. All authors contributed to creating tables, revising the manuscript, and approving the submitted version.

FUNDING

This research was supported by grants from the National Institute of Mental Health (R21MH112578), the National Institute on Drug Abuse (R34DA050256), and the National Institute of Food and Agriculture, U.S. Department of Agriculture (ILLU-793-368) to the first author.

ACKNOWLEDGMENTS

We are immensely grateful to the families who participated in this study. We also thank Kathy Sullivan, Emily Blim, Natalie Maltby, Elizabeth Mooney, Sarah Weldy, and Erin Orentas for their efforts in coordinating and conducting research visits, as well as numerous undergraduate students who assisted with data collection and observational coding.

REFERENCES

- Ainsworth, M. D. S., Blehar, M. C., Waters, E., and Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation*. Hillsdale, NJ: Erlbaum.
- Ashworth, M., Palikara, O., Burchell, E., Purser, H., Nikolla, D., and Van Herwegen, J. (2021). Online and face-to-face performance on two cognitive tasks in children with Williams Syndrome. *Front. Psychol.* 11, 1–10. doi: 10.3389/fpsyg.2020.594465
- Bayley, N. (2006a). *Bayley Scales of Infant Development: Screening Test (3rd ed.)*. London: Pearson.
- Bayley, N. (2006b). *Bayley Scales of Infant Development (3rd ed.)*. London: Harcourt Assessment, Inc.

- Braungart-Rieker, J., Garwood, M. M., Powers, B. P., and Notaro, P. C. (1998). Infant affect and affect regulation during the still-face paradigm with mothers and fathers: the role of infant characteristics and parental sensitivity. *Development. Psychol.* 34, 1428–1437. doi: 10.1037/0012-1649.34.6.1428
- Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980
- Busuito, A., and Moore, G. A. (2017). Dyadic flexibility mediates the relation between parent conflict and infants' vagal reactivity during the face-to-face still-face. *Development. Psychobiol.* 59, 449–459. doi: 10.1002/dev.21508
- Cabrera, N. J., Volling, B. L., and Barr, R. (2018). Fathers are parents, too! widening the lens on parenting for children's development. *Child Development. Perspect.* 12, 152–157. doi: 10.1111/cdep.12275
- Cleary, P. D., and Edgman-Levitan, S. (1997). Health care quality. incorporating consumer perspectives. *JAMA* 278, 1608–1612. doi: 10.1001/jama.1997.03550190072047
- Costigan, C. L., and Cox, M. J. (2001). Fathers' participation in family research: Is there a self-selection bias? *J. Fam. Psychol.* 15, 706–720. doi: 10.1037/0893-3200.15.4.706
- Davis, D. W., and Logsdon, M. C. (2011). *Maternal Sensitivity: A Scientific Foundation for Practice*. Hauppauge, NY: Nova Science Publishers.
- Garrisi, K., King, C. J., Mullin, L. J., and Gaab, N. (2020). General recommendations and guidelines for remote assessment of toddlers and children, in response to the covid-19 pandemic. *JAMA* 20,43. doi: 10.31219/osf.io/wg4ef
- Germin, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B. (2012). Is the Web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bull. Rev.* 19, 847–857. doi: 10.3758/s13423-012-0296-9
- Griffiths, S., Jarrold, C., Penton-Voak, I. S., Woods, A. T., Skinner, A. L., and Munafò, M. R. (2019). Impaired recognition of basic emotions from facial expressions in young people with Autism Spectrum Disorder: assessing the importance of expression intensity. *J. Autism Development. Disord.* 49, 2768–2778. doi: 10.1007/s10803-017-3091-7
- Gustafsson, H. C., Young, A. S., Stamos, G., Wilken, S., Brito, N. H., Thomason, M. E., et al. (2021). Innovative methods for remote assessment of neurobehavioral development. *Development. Cogn. Neurosci.* 52, 101015. doi: 10.1016/j.dcn.2021.101015
- Kelleher, B. L., Halligan, T., Witthuhn, N., Neo, W. S., Hamrick, L., and Abbeduto, L. (2020). Bringing the laboratory home: PANDABox telehealth-based assessment of neurodevelopmental risk in children. *Front. Psychol.* 11, 1–14. doi: 10.3389/fpsyg.2020.01634
- Lourenco, S. F., and Tasimi, A. (2020). No participant left behind: Conducting science during COVID-19. *Trends Cogn. Sci.* 24, 583–584. doi: 10.1016/j.tics.2020.05.003
- Maitre, N. L., Benninger, K. L., Neel, M. L., Haase, J. A., Pietruszewski, L., Levengood, K., et al. (2021). Standardized neurodevelopmental surveillance of high-risk infants using telehealth: implementation study during COVID-19. *Pediatric Qual. Safety* 6, e439. doi: 10.1097/pq9.0000000000000439
- Manning, B. L., Harpole, A., Harriott, E. M., Postolowicz, K., and Norton, E. S. (2020). Taking language samples home: feasibility, reliability, and validity of child language samples conducted remotely with video chat vs. in-person. *J. Speech, Lang. Hear. Res.* 63, 3982–3990. doi: 10.1044/2020_JSLHR-20-00202
- McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica* 22, 276–282. doi: 10.11613/BM.2012.031
- Mesman, J., van IJzendoorn, M. H., and Bakermans-Kranenburg, M. J. (2009). The many faces of the Still-Face Paradigm: A review and meta-analysis. *Development. Rev.* 29, 120–162. doi: 10.1016/j.dr.2009.02.001
- Moore, G. A., and Calkins, S. D. (2004). Infants' vagal regulation in the still-face paradigm is related to dyadic coordination of mother-infant interaction. *Development. Psychol.* 40, 1068–1108. doi: 10.1037/0012-1649.40.6.1068
- Moore, G. A., Cohn, J. F., and Campbell, S. B. (2001). Infant affective responses to mother's still face at 6 months differentially predict externalizing and internalizing behaviors at 18 months. *Develop. Psychol.* 37, 706–714. doi: 10.1037/0012-1649.37.5.706
- Pratt, M., Singer, M., Kanat-Maymon, Y., and Feldman, R. (2015). Infant negative reactivity defines the effects of parent-child synchrony on physiological and behavioral regulation of social stress. *Develop. Psychopathol.* 27, 1191–1204. doi: 10.1017/S0954579415000760
- Prime, H., Wade, M., and Browne, D. T. (2020). Risk and resilience in family well-being during the COVID-19 pandemic. *Am. Psychol.* 75, 631–643. doi: 10.1037/amp0000660
- Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., et al. (2020). Advancing developmental science via unmoderated remote research with children. *J. Cogn. Develop.* 21, 477–493. doi: 10.1080/15248372.2020.1797751
- Ruiz, M., Carrasco, M., and Holgado-Tello, F. (2019). Father involvement and children's psychological adjustment: maternal and paternal acceptance as mediators. *J. Fam. Stud.* 25, 151–169. doi: 10.1080/13229400.2016.1211549
- Scott, K., Chu, J., and Schulz, L. (2017). Lookit (Part 2): Assessing the viability of online developmental research, results from three case studies. *Open Mind* 1, 15–29. doi: 10.1162/OPMI_a_00001
- Scott, K., and Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind* 1, 4–14. doi: 10.1162/OPMI_a_00002
- Sheskin, M., and Keil, F. (2018). TheChildLab.com a Video Chat Platform for Developmental Research.
- Shin, E., Smith, C. L., and Howell, B. R. (2021). Advances in behavioral remote data collection in the home setting: Assessing the mother-infant relationship and infant's adaptive behavior via virtual visits. *Front. Psychol.* 12, 822. doi: 10.3389/fpsyg.2021.703822
- Smith-Flores, A. S., Perez, J., Zhang, M. H., and Feigenson, L. (2021). Online measures of looking and learning in infancy. *JAMA* 21,598. doi: 10.31234/osf.io/tdbnh
- Stravish, A. V., Tronick, E., Hollenstein, T., and Beeghly, M. (2013). Dyadic flexibility during the face-to-face Still-Face paradigm: a dynamic systems analysis of its temporal organization. *Infant. Behav. Dev.* 36, 432–437. doi: 10.1016/j.infbeh.2013.03.013
- Su, I., and Ceci, S. (2021). "Zoom Developmentalists": Home-based videoconferencing developmental research during COVID-19.
- Sutherland, R., Trembath, D., Hodge, A., Drevensek, S., Lee, S., Silove, N., et al. (2017). Telehealth language assessments using consumer grade equipment in rural and urban settings: Feasible, reliable and well tolerated. *J. Telemed. Telecare* 23, 106–115. doi: 10.1177/1357633X15623921
- Tran, M., Cabral, L., Patel, R., and Cusack, R. (2017). Online recruitment and testing of infants with Mechanical Turk. *J. Experiment. Child Psychol.* 156, 168–178. doi: 10.1016/j.jecp.2016.12.003
- Tronick, E., Als, H., Adamson, L., Wise, S., and Brazelton, T. B. (1978). Infants' response to entrapment between contradictory messages in face-to-face interaction. *J. Am. Acad. Child Adolesc. Psychiatry.* 17, 1–13. doi: 10.1016/S0002-7138(09)62273-1
- Tronick, E., Als, H., and Brazelton, T. B. (1980). Monadic phases: a structural descriptive analysis of infant-mother face to face interaction. *Merrill Palmer Q.* 26, 3–24. Available online at: https://www.jstor.org/stable/23084074?seq=1#metadata_info_tab_contents (accessed August 12, 2018).
- van Dijk, J. A. G. M. (2020). *The Digital Divide*. Cambridge, England: Polity.
- Yessis, J. L., Kost, R. G., Lee, L. M., Collier, B. S., and Henderson, D. K. (2012). Development of a research participants' perception survey to improve clinical research. *Clinic. Transl. Sci.* 5, 452–460. doi: 10.1111/j.1752-8062.2012.00443.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 McElwain, Hu, Li, Fisher, Baldwin and Bodway. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: frontiersin.org/about/contact



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership