frontiers | Research Topics

Remote online language assessment: Eliciting discourse from children and adults

Edited by
Natalia Gagarina, Angel Chan and Wenchun Yang

Published in
Frontiers in Communication
Frontiers in Psychology

ISSN 1664-8714
ISBN 978-2-8325-5142-4
DOI 10.3389/978-2-8325-5142-4

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

Remote online language assessment: Eliciting discourse from children and adults

Topic editors

Natalia Gagarina — Leibniz Center for General Linguistics (ZAS), Germany

Angel Chan — Hong Kong Polytechnic University, Hong Kong, SAR China

Wenchun Yang — Xi'an Jiaotong University, China

Citation

Gagarina, N., Chan, A., Yang, W., eds. (2024). *Remote online language assessment: Eliciting discourse from children and adults*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-5142-4

Table of contents

frontiers | Frontiers in Communication

TYPE Editorial
PUBLISHED 01 August 2024
DOI 10.3389/fcomm.2024.1463182

OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Wenchun Yang
✉ wenchunyang@xjtu.edu.cn
Angel Chan
✉ angel.ws.chan@polyu.edu.hk
Natalia Gagarina
✉ gagarina@leibniz-zas.de

RECEIVED 11 July 2024
ACCEPTED 16 July 2024
PUBLISHED 01 August 2024

CITATION
Yang W, Chan A and Gagarina N (2024) Editorial: Remote online language assessment: eliciting discourse from children and adults. *Front. Commun.* 9:1463182. doi: 10.3389/fcomm.2024.1463182

Editorial: Remote online language assessment: eliciting discourse from children and adults

Wenchun Yang[1]*, Angel Chan[2,3,4]* and Natalia Gagarina[5]*

[1]School of Foreign Studies, Xi'an Jiaotong University, Xi'an, China, [2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China, [3]Research Centre for Language, Cognition, and Neuroscience, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China, [4]The Hong Kong Polytechnic University—Peking University Research Centre on Chinese Linguistics, Kowloon, Hong Kong SAR, China, [5]Leibniz-Centre General Linguistics (ZAS), Berlin, Germany

KEYWORDS

remote online, language assessment, discourse context, narratives, children, adults

Editorial on the Research Topic
Remote online language assessment: eliciting discourse from children and adults

1 Introduction

Being able to collect valid data is crucial for empirical science disciplines such as linguistics, developmental psycholinguistics, clinical psycholinguistics and speech, and hearing sciences. In recent years there has been an increasing use of digital devices for remote language assessments, such as online elicitation of language samples, apps for eliciting expressive lexical abilities, online questionnaires, and other digital platforms.

The COVID-19 pandemic had affected and is still affecting many lives globally, having disrupted face-to-face, in-person language assessments, and causing many researchers to conduct their language assessments online. This shift is seen in multiple disciplines and settings, with online methods of elicitation being increasingly used not only in linguistics and other disciplines but also in clinical and educational settings.

Discourse involves verbal/written narration or exchange/conversation and linguistically not only goes beyond the sentence level but also involves language skills at different levels and their integration. Assessing an individual's competence at the discourse level is an informative indicator of one's general communication and social skills, and language and cognitive development, and could be an index of one's educational outcome. Analyzing samples of narrative discourse also allows one to examine the effects of cultural practices and properties. Given the significance of assessing the discourse competence of an individual, being able to administer the assessment via remote online means allows one to collect these informative data when there are restrictions on in-person administration of assessments.

Despite the necessity of remote language assessments and the convenience they may bring to both assessors and assesses, the potential merits, limits, and problems of remote testing have not yet been systematically explored and understood. This timely Research

Topic seeks contributions that mobilize new evidence and/or insightful and nuanced discussions to address questions such as: can we control online testing so that it is as good as face-to-face, in-person testing, and, if so, how? Do we have evaluative evidence of such practices, and if so, how robust is the evidence? What adaptations and concerns can and cannot be accommodated at the present? What opportunities are offered by recent technological advances? Are there certain conditions in which online testing works better or worse? Finally, how do differences between offline, in-person language assessments and online, remote assessments affect the results of testing?

The current Research Topic has two main foci: the first addresses the feasibility of assessing abilities at the discourse level (narrative or conversational) in both children and adults using remote online testing. Communicative competence at the discourse level has been considered an essential and ecologically valid component in language assessments of children and adults, for three key reasons: (1) this competence is crucial for an individual's everyday functioning and academic and social life, (2) it provides information about an individual's socio-cognitive and linguistic abilities, and (3) it is a versatile test of language skills at the levels of content, form, use and their integration. The second focus addresses the reliability of remote online testing in terms of comparing the results elicited via remote online assessments and in-person assessments.

We first give a general summary including an overview of the participants, languages, and methods featured in this Research Topic of papers, and then highlight the key results or significance of the specific papers. A short conclusion will close our Research Topic Introduction.

This Research Topic "*Remote online language assessment: eliciting discourse from children and adults*" intends to cover empirical articles discussing new evidence, perspective and opinion papers on issues at the conceptual-methodological interface, and methods articles presenting approaches that can offer opportunities for remote testing of discourse supported by recent technological advances. Ten papers were accepted for publication each of which has gone through the usual rigorous peer review process, and these selected papers include one perspective paper, two methods papers, and seven original research articles.

The age of participants ranged from 3 to 70 years and the number of participants per study ranged from 25 up to 4,517 participants/profiles. Five out of the seven research articles reported on bilinguals, e.g., Bosnian in the context of German, Irish-English, Mandarin-English, French-English, and two studies were dedicated to monolingual Greek and English speakers.

Four studies featured individuals with communication disorders, for example children with Autism Spectrum Disorder (Butler et al.), children with acquired reading and writing impairments (Jaecks and Jonas), adults with language impairment (Stamouli et al.), and adolescents and adults with Down Syndrome (Mattiauda et al.). A total of seven languages were featured: Bosnian, Canadian French, English, Greek, Irish, German, and Mandarin.

2 This volume

Moving onto introducing each paper in this Research Topic, we highlight each paper as follows. The perspectives paper (Jaecks and Jonas) advocated for the importance of assessing written discourse via digital means to improve social participation and digital participation for individuals with acquired reading and writing impairments and argued that remote assessment of written discourse abilities in functional communicative activities can be incorporated in teletherapy.

The two methodological papers, by Stamouli et al. and Bright et al., reported on the use of digital methods to elicit narratives from adults with(out) language impairment and from children. Stamouli et al.'s paper compared two modes of narrative elicitation methods in 10 healthy adults in a within-participants study design: remote online and in-person; and reported largely no significant differences in the narrative measures between the two elicitation methods. Bright et al. designed an app to collect story retelling samples from children. A citizen science approach was adopted to collect large samples of data and a stratified sampling framework was used to further screen participants. A total of 4,517 profiles from 599 children were collected and analyzed. Their paper demonstrated that a citizen science approach using the app is an efficient way to collect large amounts of informative research data.

The seven research papers reported on oral discourse produced by typical and atypical children, adolescents, and adults from various language backgrounds. Yang et al. examined the story-retelling skills of Mandarin-English bilingual children aged 3–6 years old ($N = 25$) using a remote method. They examined the effects of age and language experience on children's production of narrative macrostructure (the global organization of a story) and microstructure (the use of linguistic forms in the target language in a story). Their children showed comparable performance in macro-and micro- structures across the two languages. Age was significantly positively correlated with macrostructure in both languages, but no significant correlations were registered between language experience and narrative macrostructure and microstructure elements.

Burchell et al. compared the narrative and vocabulary measures collected by online and in-person assessments in two groups of children aged 7–12 years old: 127 English monolinguals and 78 French-English bilinguals. The two groups of children showed no differences between the two testing modes in both narrative discourse and receptive vocabulary measures. However, the authors reported that there are some modality differences between testing modes for the conversational and expository discourse measures.

Butler et al. examined the effect of remote natural language sampling on the interactions between parents and children with Autism Spectrum Disorder (ASD) at home. Naturalistic language samples from 90 dyads of parents and ASD children aged 4–7 years old were collected remotely when the interactions took place in the home. The range of activities and the relationship between activities and children's language levels were analyzed. The authors found no effect of the types of activities on the richness of language elicited and there was an association between the number of different activities and the child's language level.

Jažić et al. investigated the relationship between the history of language acquisition, current usage of language, and socio-economic status (SES) and case marking accuracy in 20 monolingual and 20 heritage Bosnian speakers aged 18–30 years old. They used the Multilingual Assessment Instrument for Narratives (LITMUS-MAIN) to elicit narrative discourse online. Heritage speakers showed significantly lower accuracy in case marking compared to the monolingual group. The use of Bosnian and the frequency of current usage, but not SES, were significant predictors of participants' case accuracy.

Mattiauda et al. made a first attempt to assess narrative retelling in adults with Down syndrome online using LITMUS-MAIN and compared the performance between 13 adults with Down syndrome aged 15–33 years old and a typically developing control group aged 4–10 years old. Participants with Down Syndrome were outperformed by the control group on measures of story structure, story comprehension, and lexical diversity, whereas there was no difference between the two groups in the total number of words. The authors concluded that remote online assessment of individuals with Down syndrome is feasible.

Zhou et al. reported the effects of structural similarities and differences between the languages, language input, and working memory on reference production in 4–6-year-old Mandarin-English bilingual preschoolers. They administered two stories using LITMUS-MAIN online and analyzed character introduction and reintroduction in the elicited oral discourse. These bilingual children showed a prolonged development of felicitous reference expressions and over-reliance on overt marking of definiteness in narratives. The frequency of felicitous reference expressions in the input was a significant predictor of the production of felicitous reference expressions and there was a modulating effect of working memory.

Antonijevic et al. assessed production and comprehension of narrative macrostructure in 30 adult Irish-English bilinguals online using LITMUS-MAIN. The authors found no difference in story structure, comprehension scores, and the overall number of Internal State Terms across languages. They highlighted that online assessment increases accessibility to participants, in particular, those in rural areas with low population density, whereas an unstable internet connection could limit the applicability of remote online assessment.

All contributions in this volume demonstrated that remote online language assessment of oral discourse is feasible for those children, adolescents and adults with and without language impairments examined. One benefit of using online assessment is increasing the accessibility to participants, which facilitates researchers in collecting large amount of language samples. Compared to in-person mode of assessment, remote online testing requires an environment well-equipped to support remote data collection such as a stable internet connection.

3 Future directions

With our Research Topic, we hope to be able to document new data featuring assessment of discourse competence in children and adults using remote and in-person experimental settings. We also hope to be able to suggest some directions for future research. These directions might be centered around investigations of the properties of child and adult (narrative) discourse looking for similarities across and differences within developmental trajectories. Cross-cultural research might shed light on the question of how specific cultural background factors shape discourse production and comprehension. One specific direction could involve methodological issues, e.g., development of new methods for remote elicitation of production and comprehension of discourse and its components, data collection for longitudinal and naturalistic data and multimodal data integration. The other direction is the development and validation of assessment instruments. This could include the integration of technology, for instance artificial intelligence, into the analyses of elicited discourse and the development of intervention practices. Digital tools for adaptive and personalized testing and automated scoring for assessment materials targeting discourse could also be one of the future Research Topics. Last but not the least are studies on practical issues dealing with the comparison of off-line and online assessment tools for elicitation and analyses of discourse. As remote online testing becomes more and more prevalent, one important issue is ethical and privacy considerations. Clear and robust protocols and guidelines are necessary to ensure the responsible conduct of research on remote online language assessment.

Author contributions

WY: Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization. AC: Writing – review & editing, Writing – original draft, Conceptualization, Funding acquisition. NG: Writing – review & editing, Writing – original draft, Conceptualization, Funding acquisition.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the Fundamental Research Funds for the Central Universities (SK2024025) and a research stipend by the Fritz Thyssen Foundation (40.20.0.002SL) to WY, and a research grant (P0014049; G-YW4G; Chief supervisor: AC; Co-supervisor: NG), awarded by the Research Grants Council General Research Fund, Hong Kong, to AC and NG.

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

frontiers | Frontiers in Communication

PERSPECTIVE
published: 03 May 2022
doi: 10.3389/fcomm.2022.798143

Digital Assessment of Acquired Alexia and Agraphia: On the Relevance of Written Discourse

Petra Jaecks[1]* *and Kristina Jonas*[2]

[1] Faculty of Linguistics and Literary Studies, Bielefeld University, Bielefeld, Germany, [2] Department of Special Education and Rehabilitation, Faculty of Human Sciences, University of Cologne, Cologne, Germany

The digital revolution has created challenges as well as opportunities for people with acquired reading (= alexia) and writing (= agraphia) impairments. Although it is difficult to validly assess written discourse, it is imperative that people with alexia and agraphia (PwAA) receive reliable diagnostics for the following reasons: (1) discourse in written and oral forms is highly relevant to daily interaction and participation, but there are no established tests or diagnostic procedures to assess written discourse; (2) reliable diagnostic measures are a prerequisite for any language rehabilitation, especially for the complex skills needed for written discourse; and (3) the continuing trend in digitalization offers new opportunities for easily collecting and assessing written discourse via digital means. In our manuscript, we highlight the relevance of written discourse for social participation and in the digital world and argue that in order to improve social participation in general and digital participation in particular for PwAA, remote assessment of written discourse abilities can be the basis for speech and language therapy treatment focused on communicative abilities.

Keywords: alexia, agraphia, digital, assessment, diagnostic, discourse, aphasia

OPEN ACCESS

Edited by:
Wenchun Yang,
Leibniz-Centre for General Linguistics (ZAS), Germany

Reviewed by:
Jeremy Purcell,
University of Maryland, College Park, United States
Giorgio Arcara,
San Camillo Hospital (IRCCS), Italy

***Correspondence:**
Petra Jaecks
petra.jaecks@uni-bielefeld.de

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *19 October 2021*
Accepted: *12 April 2022*
Published: *03 May 2022*

Citation:
Jaecks P and Jonas K (2022) Digital Assessment of Acquired Alexia and Agraphia: On the Relevance of Written Discourse. Front. Commun. 7:798143. doi: 10.3389/fcomm.2022.798143

INTRODUCTION

The rapid increase in the use of digital technologies in recent years—and the accelerated development of such technologies in response to the challenges posed by the COVID-19 pandemic—has changed society (cf. Berner et al., 2020; United Nations, 2021), resulting in specific challenges for certain groups of people while also allowing a set of new opportunities. These challenges and opportunities arise particularly for people with acquired reading impairments (= alexia) and writing impairments (= agraphia; people with alexia and agraphia = PwAA).

Alexia and agraphia are common symptoms of acquired language disorders after injury has occurred to the brain, and in 60% of cases they occur in addition to general impairments in language production and reception (= aphasia; see Brookshire et al., 2014; Rapcsak and Beeson, 2015; Riley et al., 2015). The classification of alexia and agraphia differentiates between central and peripheral disorders of written language processing, i.e., "*classified as "central" (or linguistic) when it is generated at a level that affects spelling and as "peripheral" when the spelling is correctly generated but the peripheral procedures are not correctly activated*" (Silveri et al., 2007, p. 179). Furthermore, it is relevant whether deficits are present in the context of aphasia or as pure alexia or pure agraphia (e.g., Rapcsak and Beeson, 2015; Riley et al., 2015; Schumacher et al., 2020; Sheppard and Sebastian, 2020). Typical symptoms of alexia and agraphia include phonemic, semantic, and morphemic

paralexia as well as graphemic, semantic and formal paragraphia; regularization errors[1] also occur. These symptoms are observed in reading and writing skills at the word and sentence levels as well as in written discourse (e.g., Leff and Behrmann, 2008; Tiu and Carter, 2021).

Discourse skills in general, be they written or oral, take on a special significance in social interaction: they serve to achieve a (communicative) goal as well as exchange/communicate information (cf. Armstrong et al., 2012; Dipper and Pritchard, 2017). Pickering and Garrod's (2004) Alignment Theory is a communication theory that focuses extensively on conversational success. This theory seems particularly suitable as a basis for the analysis of written discourse, since it explicitly includes the underlying levels of language processing, e.g., the lexicon and syntax. Pickering and Garrod assume that an alignment of, for example, the syntactic structure contributes to an alignment of the situation model and thus facilitates conversation. Foltz et al. (2015) were able to show that alignment processes can also be found in written interaction (see also Michel and Cappellini, 2019). Kim et al. (2019) found even more alignment in written vs. oral communication. Since alignment can facilitate the processing of linguistic (written) utterances, Pickering and Garrod's (2004) alignment theory is particularly useful for the analysis of impaired (written) discourse and should be taken into account in the development of diagnostic procedures as a whole as well as for the concrete creation of the task and items.

It is important to note that the processes underlying reading and writing at the word level (cf. e.g., Caramazza and Miceli, 1990; Miceli and Capasso, 2006) are also relevant at the discourse level. As in the context of the Alignment Theory mentioned above, inter-level influences must be assumed. However, the focus of this perspective paper lies on written discourse and its remote assessment possibilities.

OUR PERSPECTIVE

In thinking about (written) discourse, we need to distinguish between different forms, i.e., conversational, procedural, persuasive, personal, descriptive, expository, and narrative discourse (e.g., Dipper and Pritchard, 2017; Zanichelli et al., 2020).

Notably, these different forms of (written) discourse differ in terms of their relevance to everyday life, and discourse types that have a high everyday relevance in oral discourse are not necessarily as relevant for everyday written discourse (cf. Grotlüschen et al., 2020).

Moreover, not all forms of oral discourse can be found in written discourse. Narratives, such as those generated by describing picture stories or personal experiences, or semi-directed interviews (cf. Zanichelli et al., 2020), are not very frequent in written form. Other forms of discourse, such as conversations, are becoming more common in written form, e.g., email correspondence or chat communication (e.g., Dietz et al., 2011; Grotlüschen et al., 2020).

It is therefore not very useful to simply transfer established methods for eliciting oral discourse to written language. Instead, there is a need to develop and test tasks that are suitable for eliciting and analyzing written discourse relevant to everyday life (e.g., Steel and Togher, 2019; Johansson-Malmeling et al., 2021).

Although the production of narratives is the most frequently described and studied variant of discourse (e.g., Behrns et al., 2010; Bryant et al., 2017; Steel and Togher, 2019), especially for people with language disorders, the analysis of narratives reveals unanswered questions and challenges.

First, it is not always clear which characteristics of spoken/written language are a result of an idiomatic style and at what point certain conspicuous features or peculiarities, as for example elliptic utterances in chat conversations, should be interpreted as pathological, especially in discourse production (e.g., Obermeyer and Edmonds, 2018; Schweiger, 2018).

Another reason is there are still few formal specifications, and open questions remain concerning the analysis of written discourse competencies: what is rated—the writing process, including self-corrections, or the final written text? (e.g., Johansson-Malmeling et al., 2021). What about the time needed to write or understand an answer or question? How is the influence of the conversation partner included in an analysis of discourse competence?

A further reason underlying the difficulty of analyzing written discourse is there is only limited information about the cognitive prerequisites and processes needed for written discourse, especially in PwAA (e.g., Behrns et al., 2010). There are written task formats (e.g., email facilitated interviews, cf. Egan et al., 2006) where the process of completing the task, which is relevant for a valid assessment and goal-oriented diagnosis, cannot be observed directly (e.g., Johansson-Malmeling et al., 2021). But this type of asynchronous conversation offers the possibility of more time for comprehension and formulation, and it is likely that less working memory capacity is required to complete the task. This again emphasizes the complexity involved in assessing what resources were used to solve a task and what support was available.

Further difficulties concern the even greater influence of education levels on written language competences (e.g., Zanichelli et al., 2020) and organizational and practical reasons in clinical settings such as time limits or data protection issues (e.g., Bryant et al., 2017; Steel and Togher, 2019; Obermeyer et al., 2021).

The factors listed above can be roughly assigned to four categories: (a) cognitive and linguistic questions, e.g., the theories and processes underlying "normal" written discourse, (b) situational and contextual challenges, e.g., the situation and the participants' personality, competences, or motivation, (c) further technological difficulties, e.g., technical equipment and digital methods, and (d) the necessity of having ecological validity.

Although it is challenging to validly assess written discourse, we call for the development of reliable writing-based diagnostics for people with alexia and agraphia (PwAA). In this perspective paper, we will first clarify the need for and advantages of a

[1]Regularization errors are present when, "an irregularly spelled word is mispronounced by incorrect application of regular spelling-sound correspondences (e.g., reading plaid as "played"), indicating over-reliance on sublexical grapheme-phoneme correspondences" (Binder et al., 2016, p. 1).

diagnostic for written discourse and then propose one example. Several facts justify our position.

First, written discourse is as relevant for social participation as oral discourse is in everyday communication (e.g., Dietz et al., 2011; Obermeyer et al., 2021), yet there are hardly any tests that specifically assess written discourse (cf. Bryant et al., 2017; Rohde et al., 2018; Steel and Togher, 2019). The analysis of spontaneous speech or oral discourse is a frequent and important component of aphasia diagnostics (cf. Stark et al., 2021). The Aachen Aphasia Test (AAT, Huber et al., 1983) assesses spontaneous speech based on a semi-standardized interview using six 6-level scales (communication, articulation, automated speech, semantics, phonology, syntax). In the Western Aphasia Battery (WAB, Kertesz, 2007) discourse competence is looked at in two subtests. First, six personal questions must be answered, followed by a picture description. The evaluation of both language samples is carried out on two levels ("Information content" and "Fluency, grammatical competence, and paraphasias"). The production and reception of written discourse, in contrast, usually play a minor role in traditional standard procedures (e.g., Bryant et al., 2017). In the AAT, reading and writing are only assessed at the word and sentence level. The WAB goes a step further and includes a test section with a written picture description. Written discourse, be it digital or analog, is meaningfully impaired in alexia and agraphia (e.g., Mortensen, 2005; Behrns et al., 2010; Johansson-Malmeling et al., 2021). Because the everyday linguistic-communicative competences necessary for social participation and a decent (communication-related) quality of life (Neumann et al., 2019) are not limited to oral performance, written language skills must always be taken into account as well (e.g., Mortensen, 2005; Dietz et al., 2011).

Second, the analysis of written language must go beyond the word and sentence level and include discourse. The major deficits in written discourse in PwAA reinforce the need for reliable diagnostics at the discourse level. Tests that do check reading and writing skills in acquired alexia or agraphia often refer only to the word or sentence level. A typical example is the newly established procedure DYMO (DYslexien MOdellorientiert, Schumacher et al., 2020), which was developed from the two-route model of Coltheart et al. (1993) to examine acquired reading disorders in German. In several subtests, many components of reading are tested, but only at the word level. In order to be able to address written language competences at all relevant levels from function to activity, and hence from word to discourse level, a well-founded diagnostic is necessary [e.g., National Stroke Foundation, 2010; Rohde et al., 2018]. Only in this way can relevant dysfunctions be identified and restrictions in participation be reliably detected. This is one of the essential prerequisites for individual therapy planning, which also includes the personal resources of the person concerned (Gerhards et al., 2022).

Third, our society is being shaped by the ongoing process of digitalization, and as new technologies become more and more important, we experience an increasingly written environment. Despite this, the degree of digitalization is only beginning in speech and language therapy with PwAA (cf. Bilda et al., 2016; Weidner and Lowman, 2020). Nevertheless, we should use the advantages and possibilities of digitalization for complex diagnostic issues (e.g., Jonas and Jaecks, 2021), as in the case of written discourse in PwAA. In an increasingly digital environment (e.g., when people use messaging services and online portals to get in touch with each other), written discourse plays an ever more important role. Similar to the consequences of functional illiteracy on social participation (Cree et al., 2012; Vágvölgyi et al., 2016), there are clear disadvantages and difficulties for PwAA in a "written world," no matter whether digital or analog.

Here we briefly introduce two possible test situations that enable the analysis of written discourse. One of the challenges of everyday life is communicating with virtual agents or bots on the Internet. A conceivable diagnostic scenario close to everyday life, for example, is a chat in a complaint or customer care portal. EVA Park (cf. Marshall et al., 2020) is an online virtual world designed for people with acquired language disorders. Although it was not developed for diagnostic purposes, the virtual environment contains various therapeutic tasks and group session opportunities. An everyday communication test scenario could be programmed within this platform, e.g., the PwAA is tasked with buying a ticket and answering questions presented by a virtual agent. This type of assessment involves high ecological validity and aspects of social participation.

A second everyday scenario that also requires written discourse is communicating with family or friends via instant messaging services (see Overlach et al., 2020 for an example of therapeutic use). A diagnostic task can also be set here, such as negotiating and agreeing on an appointment time and place with one or more people. Both scenarios can be transferred directly from real life to the diagnostic situation.

As in everyday life, the examiner communicates with the PwAA via the medium, i.e., the virtual world, the messenger service or a specific website, and all reactions are saved and subsequently available for analysis. Depending on the technical conditions, it may also be possible to automate the interaction—and thus the diagnostic procedure—using adaptive algorithms.

The advantage of an automated analysis would also be that "normal" idiomatic aspects of an individual PwAA could be better identified and contrasted with pathological parts on the basis of big data analysis [see for example (Savoy, 2020) for advanced models for stylometric applications].

However, the strict data protection regulations, especially with regard to PwAA, are a challenge. Since a large amount of data is needed as a basis for standardization studies, automated analyses and the recognition of language peculiarities, creating an underlying database is correspondingly complex and will take time. Nevertheless, this approach to digital diagnostics is very promising and should be pursued further (e.g., Kohlschein et al., 2018; Torre et al., 2021).

While over recent years researchers and practitioners have developed concepts for telemedical therapy, there is almost no evidence of remote assessment of language disorders following cerebrovascular diseases (e.g., Weidner and Lowman, 2020), including for the diagnosis of alexia and agraphia (Jaecks and Jonas, 2021). However, given that digital written discourse is technically easy to collect, i.e., via remote assessment, virtual

reality settings, and other new technologies, the analysis of written discourse in PwAA can benefit from the advantages of digitalization.

Written discourse skills are much more important in the digital world than oral discourse skills. Reduced (digital) participation caused by agraphia and alexia leads to difficulties in daily communicative activities, and in turn to restrictions in all important areas of life (self-determination, educational and vocational qualifications, social contacts, etc.; cf. Grotlüschen et al., 2020; Vishal, 2021).

CONSEQUENCES

The long-term goal must therefore be the remote assessment of written discourse, based on the concept of the ICF (WHO, 2001). Moreover, communicative activities involving written discourse have to be reliably recorded. The type of written discourse and its specific relevance to everyday life, as well as the possibility of drawing conclusions for therapy, are the factors that determine the choice of survey methods. This applies to the development of the diagnostic procedure in general as well as to the use of diagnostics in individual patients. We require a digital diagnostic procedure for written discourse that can be used as an unconventional remote screening tool for PwAA following cerebrovascular incidents and allows prompt and direct access to telemedical rehabilitation, which is essential for social participation.

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/s.

AUTHOR CONTRIBUTIONS

PJ and KJ contributed equally to the conceptualization and writing of the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Armstrong, E., Bryant, L., Ferguson, A., and Simmons-Mackie, N. (2012). "Approaches to assessment and treatment of everyday talk in aphasia," in *Aphasia and Related Neurogenic Communication Disorders*, *2nd Edn,* eds I. Papathanasiou and P. Coppens (Jones and Barlett Learning), 269–285.

Behrns, I., Ahlsén, E., and Wengelin, A. (2010). Aphasia and text writing. *Int. J. Lang. Commun. Disord.* 45, 230–243. doi: 10.3109/13682820902936425

Berner, F., Endter, C., and Hagen, C. (2020). *Ältere Menschen und Digitalisierung: Erkenntnisse und Empfehlungen des Achten Altersberichts*. Bundesministerium für Familie, Senioren, Frauen und Jugend. Available online at: https://www.bmfsfj.de/blob/jump/159704/achter-altersbericht-aeltere-menschen-und-digitalisierung-data.pdf (accessed April 04, 2022).

Bilda, K., Mühlhaus, J., and Ritterfeld, U. (2016). *Neue Technologien in der Sprachtherapie*. Thieme. doi: 10.1055/b-004-129736

Binder, J. R., Pillay, S. B., Humphries, C. J., Gross, W. L., Graves, W. W., and Book, D. S. (2016). Surface errors without semantic impairment in acquired dyslexia: a voxel-based lesion-symptom mapping study. *Brain* 139(Pt 5), 1517–1526. doi: 10.1093/brain/aww029

Brookshire, C. E., Wilson, J. P., Nadeau, S. E., Gonzalez Rothi, L. J., and Kendall, D. L. (2014). Frequency, nature, and predictors of alexia in a convenience sample of individuals with chronic aphasia. *Aphasiology* 28, 1464–1480. doi: 10.1080/02687038.2014.945389

Bryant, L., Spencer, E., and Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology* 31, 1105–1126. doi: 10.1080/02687038.2016.1239013

Caramazza, A., and Miceli, G. (1990). The structure of graphemic representations. *Cognition*. 37:243–97.

Coltheart, M., Curtis, B., Atkins, P., and Haller, M. (1993). Models of reading aloud: dual-route and parallel-distributed-processing approaches. *Psychol. Rev.* 100:589. doi: 10.1037/0033-295X.100.4.589

Cree, A., Kay, A., and Steward, J. (2012). *The Economic and Social Cost of Illiteracy: A Snapshot of Illiteracy in a Global Context*. Final report from the World Literacy Foundation. World Literacy Foundation.

Dietz, A., Ball, A., and Griffith, J. (2011). Reading and writing with aphasia in the 21st century: technological applications of supported reading comprehension and written expression. *Top. Stroke Rehabil.* 18, 758–769. doi: 10.1310/tsr1806-758

Dipper, L. T., and Pritchard, M. (2017). "Discourse: assessment and therapy," in *Advances in Speech-Language Pathology* ed F. D. M. Fernandes (London: Intechopen), 3–23. doi: 10.5772/intechopen.69894

Egan, J., Chenoweth, L., and McAuliffe, D. (2006). Email-facilitated qualitative interviews with traumatic brain injury survivors: a new and accessible method. *Brain Injury* 20, 1283–1294. doi: 10.1080/02699050601049692

Foltz, A., Gaspers, J., Meyer, C., Thiele, K., Cimiano, P., and Stenneken, P. (2015). Temporal effects of alignment in text-based, task-oriented discourse. *Discour. Process.* 52, 609–641. doi: 10.1080/0163853X.2014.977696

Gerhards, L., Quinting, J., and Jonas, K. (2022). "Interpretation of results of speech language examination," in *European Manual of Medicine. Phoniatrics 2 - Speech and Speech Fluency Disorders - Literacy Development Disorders - Acquired Motor Speech and Language Disorders - Dysphagia*, eds A. am Zehnhoff-Dinnesen, A. Schindler, M.-C. Monfrais-Pfauwadel, K. Neumann, J. Sopko, and P. Zorowka (Berlin: Springer Nature).

Grotlüschen, A., Buddeberg, K., Dutz, G., Heilmann, L., and Stammer, C. H. (2020). *Hauptergebnisse und Einordnung zur LEO-Studie 2018 - Leben mit Geringer Literalität,* (Bielefeld: Bertelsmann) 14–62.

Huber, W., Poeck, K., and Weniger, D. (1983). *Aachener Aphasie Test (AAT)*. Göttingen: Hogrefe.

Jaecks, P., and Jonas, K. (2021). Digitalisierung in der Diagnostik und therapie von Schriftsprachstörungen. *Sprachtherapie Aktuell* 8:e2021–43.

Johansson-Malmeling, C., Hartelius, L., and Wengelin, Å. (2021). Written text production and its relationship to writing processes and spelling ability in persons with post-stroke aphasia aphasia. *Aphasiology* 35, 615–632. doi: 10.1080/02687038.2020.1712585

Jonas, K., and Jaecks, P. (2021). "Digitale diagnostik: innovative wege für die sprachtherapie," in *Spektrum Patholinguistik Band 14 Schwerpunktthema: Klick für Klick: Schritte in der digitalen Sprachtherapie*, eds T. Fritzsche, S. Breitenstein, H. Wunderlich, and L. Ferchland (Potsdam: Universitätsverlag Potsdam), 1–29.

Kertesz, A. (2007). *Western Aphasia Battery-Revised*. San Antonio, TX: The Psychological Corporation. doi: 10.1037/t15168-000

Kim, Y., Jung, Y., and Skalicky, S. (2019). Linguistic alignment, learner characteristics, and the production of stranded prepositions in relative clauses: comparing FTF and SCMC contexts. *Stud. Sec. Lang. Acquis.* 41, 937–969. doi: 10.1017/S0272263119000093

Kohlschein, C., Klischies, D., Meisen, T., Schuller, B. W., and Werner, C. J. (2018). "Automatic processing of clinical aphasia data collected during diagnosis sessions: challenges and prospects," in *Proceedings of Workshop RaPID-2 Int. Conf. Lang. Resour. Eval. (LREC)* (Miyazaki), 11–18.

Leff, A. P., and Behrmann, M. (2008). Treatment of reading impairment after stroke. *Curr. Opin. Neurol.* 21, 644–648. doi: 10.1097/WCO.0b013e3283168dc7

Marshall, J., Devane, N., Talbot, R., Caute, A., Cruice, M., Hilari, K., et al. (2020). A randomised trial of social support group intervention for people with aphasia: a Novel application of virtual reality. *PLoS ONE* 15:e0239715. doi: 10.1371/journal.pone.0239715

Miceli, G., and Capasso, R. (2006). Spelling and dysgraphia. *Cognit Neuropsych*, 23, 110–134.

Michel, M., and Cappellini, M. (2019). Alignment during synchronous video versus written chat L2 interactions: a methodological exploration. *Annu. Rev. Appl. Linguist.* 39, 189–216. doi: 10.1017/S0267190519000072

Mortensen, L. (2005). Written discourse and acquired brain impairment: evaluation of structural and semantic features of personal letters from a Systemic Functional Linguistic perspective. *Clin. Linguist. Phonet.* 19, 227–247. doi: 10.1080/02699200410001698652

National Stroke Foundation (NSF) (2010). *Clinical Guidelines for Stroke Management.* National Stroke Foundation.

Neumann, S., Quinting, J., Rosenkranz, A., de Beer, C., Jonas, K., and Stenneken, P. (2019). Quality of life in adults with neurogenic speech-language-communication difficulties: a systematic review of existing measures. *J. Commun. Disord.* 79, 24–45. doi: 10.1016/j.jcomdis.2019.01.003

Obermeyer, J. A., and Edmonds, L. A. (2018). Attentive reading with constrained summarization adapted to address written discourse in people with mild aphasia. *Am. J. Speech Lang. Pathol.* 27, 392–405. doi: 10.1044/2017_AJSLP-16-0200

Obermeyer, J. A., Rogalski, Y., and Edmonds, L. A. (2021). Attentive reading with constrained summarization-written, a multi-modality discourse-level treatment for mild aphasia. *Aphasiology* 35, 100–125. doi: 10.1080/02687038.2019.1686743

Overlach, F., Lürmann, N., and Bauer, A. (2020). WhatsApp in der aphasietherapie-the use of whatsapp in aphasia therapy. *Logos* 28, 253–264.

Pickering, M. J., and Garrod, S. (2004). The interactive-alignment model: developments and refinements. *Behavioral Brain Sci*, 27, 212–225.

Rapcsak, S. Z., and Beeson, P. M. (2015). "Neuroanatomical correlates of spelling and writing," in *The Handbook of Adult Language Disorders*, ed A. E. Hillis (New York, NY: Psychology Press), 87–116.

Riley, E. A., Brookshire, C. E., and Kendall, D. L. (2015). "Acquired alexias: mechanisms of reading," in *The Oxford Handbook of Aphasia and Language Disorders*, eds A. M. Raymer and L. J. Gonzales Rothi (Oxford University Press), 215–240. doi: 10.1093/oxfordhb/9780199772391.013.12

Rohde, A., Worrall, L., Godecke, E., O'Halloran, R., Farrell, A., and Massey, M. (2018). Diagnosis of aphasia in stroke populations: a systematic review of language tests. *PLoS ONE* 13:e0194143. doi: 10.1371/journal.pone.0194143

Savoy, J. (2020). *Machine Learning Methods for Stylometry*. Cham: Springer. doi: 10.1007/978-3-030-53360-1

Schumacher, R., Ablinger, I., and Burchert, F. (2020). *DYMO*. Hofheim: Nat-Verlag.

Schweiger, W. (2018). "Online-nutzung und individueller schreibstil-20 jahre später," in *Kumulierte Evidenzen* eds P. Rössler, and C. Rossmann (Wiesbaden: Springer VS), 69–90. doi: 10.1007/978-3-658-18859-7_4

Sheppard, S. M., and Sebastian, R. (2020). Diagnosing and managing post-stroke aphasia. *Expert Rev. Neurotherap.* 21, 221–234. doi: 10.1080/14737175.2020.1855976

Silveri, C. M., Corda, F., and Di Nardo, M. (2007). Central and peripheral aspects of writing disorders in Alzheimer's disease. *J. Clin. Exp. Neuropsychol.* 29, 179–186. doi: 10.1080/13803390600611351

Stark, B. C., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., et al. (2021). Spoken discourse assessment and analysis in aphasia: an international survey of current practices. *J. Speech Lang. Hear. Res.* 66, 4366–4389. doi: 10.1044/2021_JSLHR-20-00708

Steel, J., and Togher, L. (2019). Social communication assessment after traumatic brain injury: a narrative review of innovations in pragmatic and discourse assessment methods. *Brain Injury* 33, 48–61. doi: 10.1080/02699052.2018.1531304

Tiu, J. B., and Carter, A. R. (2021). *Agraphia*. Treasure Island, FL: StatPearls Publishing.

Torre, I. G., Romero, M., and Álvarez, A. (2021). Improving aphasic speech recognition by using novel semi-supervised learning methods on aphasiabank for English and Spanish. *Appl. Sci.* 11:8872. doi: 10.3390/app11198872

United Nations (2021). *Department of Economic and Social Affairs.* United Nations. Available online at: https://publicadministration.un.org/en/ict4d (accessed April 04, 2022).

Vágvölgyi, R., Coldea, A., Dresler, T., and Schrader, J. (2016). A review about functional illiteracy: definition, cognitive, linguistic, and numerical aspects. *Front. Psychol.* 7:1617. doi: 10.3389/fpsyg.2016.01617

Vishal, M. V. (2021). The digital-elderly: conceptualizing ageing in the digital era-2030-2100. *Indian J. Gerontol.* 35, 556–568.

Weidner, K., and Lowman, J. (2020). Telepractice for adult speech-language pathology services: a systematic review. *Perspect. ASHA Spec. Interest Groups* 5, 326–338. doi: 10.1044/2019_PERSP-19-00146

WHO (2001). *The International Classification of Functioning, Disability and Health - ICF.* World Health Organization.

Zanichelli, L., Fonseca, R. P., and Ortiz, K. Z. (2020). Influence of age and schooling in written discourse of healthy adults. *Psicologia* 33, 1–8. doi: 10.1186/s41155-020-00148-7

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.

frontiers | Frontiers in Communication

ORIGINAL RESEARCH
published: 10 May 2022
doi: 10.3389/fcomm.2022.820564

Remote Natural Language Sampling of Parents and Children With Autism Spectrum Disorder: Role of Activity and Language Level

Lindsay K. Butler[1]*, *Chelsea La Valle*[1], *Sophie Schwartz*[1], *Joseph B. Palana*[1], *Cerelia Liu*[1], *Natalie Peterman*[1], *Lue Shen*[2] *and Helen Tager-Flusberg*[1]

[1] Department of Psychological and Brain Sciences, Boston University, Boston, MA, United States, [2] Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA, United States

Natural language sampling (NLS) is a common methodology in research and clinical practice used to evaluate a child's spontaneous spoken language in a naturalistic context. Autism spectrum disorder (ASD) is a complex neurodevelopmental condition that results in heterogeneous language profiles. NLS has emerged as a useful method for better understanding language use and development in this population. Prior work has examined the effects that contexts (e.g., home, lab) and conversational partners (e.g., examiner, parent) have on childrens language production, but less is known about remote collection of interactions between parents and children with ASD at home. Increasing our understanding of in-home remote NLS with children with ASD will improve naturalistic approaches to language assessment in children with ASD. We analyzed natural language samples of 90 dyads of parents and four- to seven-year old children with ASD collected remotely in the home using items and activities from the family's own home. The 15-min parent-child interactions were transcribed and analyzed for the child's language level measured by the number of different words. We examined the range of activities and the relationship between activities and the child's language level. We found that in-home parent-child activities fell into 13 descriptive categories, but we found no significant difference in child's language level (measured by the mean number of different words) across activities. We found that dyads involving children with higher language levels engaged in significantly fewer different activities compared to children with lower language levels. We found no difference in the number of different words elicited in the five most frequent activities in our sample. These results support the feasibility of remote in-home language sampling. While the types of activities that parent-child dyads engaged in did not affect the richness of language elicited, the number of different activities was associated with the child's language level. Allowing parents to steer children with lower language levels toward more different activities may allow children with lower language to more fully demonstrate their spoken language abilities.

Keywords: autism spectrum disorder, natural language sampling, remote assessment, language, parent-child activities

OPEN ACCESS

Edited by:
Angel Chan,
Hong Kong Polytechnic University,
Hong Kong SAR, China

Reviewed by:
Angela T. Morgan,
Royal Children's Hospital, Australia
Laura M. Morett,
University of Alabama, United States

***Correspondence:**
Lindsay K. Butler
lbutlert@bu.edu

Specialty section:
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

Received: *23 November 2021*
Accepted: *04 April 2022*
Published: *10 May 2022*

Citation:
Butler LK, La Valle C, Schwartz S, Palana JB, Liu C, Peterman N, Shen L and Tager-Flusberg H (2022) Remote Natural Language Sampling of Parents and Children With Autism Spectrum Disorder: Role of Activity and Language Level. Front. Commun. 7:820564. doi: 10.3389/fcomm.2022.820564

1. INTRODUCTION

Natural language sampling (NLS) is a common methodology in research and clinical practice used to evaluate a child's spontaneous spoken language in a naturalistic context. It provides a more naturalistic and representative sample of a child's language use than standardized assessments (see e.g., Evans and Craig, 1992; Costanza-Smith, 2010; Sanchez et al., 2020). NLS was traditionally carried out in the research lab or clinic, but with 86.6% of families in the U.S. having smartphones or devices with internet access in the home (American Communities Survey, 2019), remote in-home NLS is feasible. With the COVID-19 pandemic, remote NLS in research and clinical practice has become a necessary tool.

For children with autism spectrum disorder (ASD), NLS has emerged as a particularly useful method of language assessment with children with ASD (Tager-Flusberg et al., 2009; Barokova and Tager-Flusberg, 2018). While some children with ASD show standardized assessment scores within one standard deviation of the mean, 30% of children with ASD are minimally or low verbal (MLV) and remain so past the age of five despite access to early and quality interventions (Tager-Flusberg and Kasari, 2013). Other children with ASD fall between verbally fluent and minimally verbal. NLS can be analyzed for a range of different language features (Miller, 1981) to assess within language heterogeneity, a salient characteristic of ASD (Barokova and Tager-Flusberg, 2018).

Previous studies suggest that children with ASD, especially those who are MLV, demonstrate their best abilities in naturalistic contexts (Tager-Flusberg and Kasari, 2013). Given that naturalistic assessment is a more optimal approach for MLV individuals with ASD, researchers have encouraged parents of children with ASD to collect NLS at home. Barokova et al. (2020)'s study included parents of MLV children with ASD who used NLS in the home using a semi-structured protocol. The researchers then compared NLS collected by researchers in the lab to those collected by parents in the home using the same protocol. They found that MLV children produced an average of seven more utterances, took two more conversational turns, and produced around 1.5 more different words during a 20-min NLS with parents compared to with an examiner. Similarly, Kover et al. (2014) found that young children with ASD produced more utterances and showed better structural and pragmatic language skills in a play-based context with a parent compared to the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2012), a semi-structured diagnostic autism assessment administered by an examiner which is commonly used as a language sampling context. Other studies have also reported on the quality and quantity of language elicited in naturalistic and home environments (see e.g., Burgess et al., 2013; Gladfelter and Van Zuiden, 2020; Hilvert et al., 2020). These findings support the influential role parents play at eliciting language from their children that is representative of their child's actual expressive language abilities and highlights the potential of remote collection of a NLS by parents at home. Given the shift toward naturalistic parent-mediated interventions for young children with ASD, remote in-home assessment using materials and everyday in-home activities will increase the proximity of individualized assessment and intervention that generalizes across people and contexts (see e.g., Schreibman et al., 2015; Bentenuto et al., 2016).

Given the COVID-19 pandemic, remote data collection has allowed researchers to continue their work when in-person collection of NLS is not feasible. Recent work supports the feasibility of remote methods for examining childrens language production. Manning et al. (2020) compared language samples from neurotypical children during parent-child play collected in the laboratory to parent-child play collected via video chat in the home. They found in-person samples and remote samples did not differ significantly in the number of usable samples or in the percent of intelligible utterances. Similarly, they found no significant differences in child speech and language characteristics (including mean length of utterance, type-token ratio, number of different words, grammatical errors/omissions, and child speech intelligibility) between in-person and remote samples. Furthermore, they investigated transcription reliability through a blinded comparison of 25% of the remote and in lab samples by dividing the number of matching words, morphemes, and codes between the two transcripts by the total words/morphemes/codes, and did not differ significantly for samples collected in-person vs. remotely. They reported high transcription reliability between in-lab and at-home language samples (M = 88.59%; Range = 82–98%).

Although prior work has examined the effects of different language sampling contexts (e.g., home, lab) and conversational partners (e.g., examiner, parent) on childrens language production, less is known about remote collection of interactions between parents and children with ASD at home, particularly when parents are given open-ended elicitation instructions and use items and materials they have in the home. Our goal is to understand open-ended NLS with parents and items in the child's own home, including the type and number of activities and the relationship to the child's language level. Exploring specific activities during parent-child interactions in the home can allow for a richer and ecologically valid assessment of childrens spoken language abilities compared to standardized assessments with unfamiliar adults in a lab or clinic setting (see e.g., Costanza-Smith, 2010). Such work can provide insights into the role of in-home NLS in the assessment of language for children with ASD and inform individualized parent-mediated interventions.

Autism assessment practices have evolved significantly over the past three decades (Rosen et al., 2021). Due to the COVID-19 pandemic, the Autism Science Foundation convened a panel of senior clinicians, researchers and a professional parent-leader to re-envision the autism assessment process in light of pandemic-related experiences. The COVID-19 pandemic presents a unique opportunity to step back and review ASD assessment with the goal of developing accessible, flexible and sustainable practices (Zwaigenbaum et al., 2021). The development and adaptation of remote assessment tools can meet the demands of the pandemic and also provide an opportunity to refine assessment methods so that they are more equitable across demographic characteristics (e.g., race,

ethnicity, sex, gender), as well as feasible across cultures (Franz et al., 2017; Dash et al., 2021).

1.1. Goals of the Current Study

The overall goal of this study is to understand NLS via naturalistic interactions between parents and children with ASD collected remotely in the home using items, materials and activities from the family's own home. It is essential to understand naturalistic, open-ended remote language sampling because it can open the window to accessible and equitable practices for language assessment, particularly for children with ASD who are MLV.

The current study seeks to answer the following specific questions:

1. What activities do parents choose to promote spoken language with their children in a remote context?
2. Does type of activity or number of activities depend on child language level (measured by number of different words)?
3. Do different activity types elicit more language from children (measured by number of different words)?

2. METHODS

Study data were collected and managed using REDCap (Research Electronic Data Capture) electronic data capture tools (Harris et al., 2009, 2019) hosted at Boston University. REDCap is a secure, HIPAA compliant, web-based software platform for research studies, providing an interface for validated data capture and data manipulation and export.

2.1. Participants

We enrolled a total of 105 families, of which 13 were recruited from social media advertising and 92 were recruited through the Simons Foundation Powering Autism Research for Knowledge (SPARK) research match registry (Feliciano et al., 2018). SPARK is a national ASD genotyping project that recruits families from 31 U.S. academic medical centers with over 70,000 families enrolled. Once families enroll, they are offered the opportunity to continue hearing about and engaging in prospective research opportunities through their online research registry. SPARK has been shown to have high validity for autism diagnosis. Based on two different methods of confirming ASD diagnosis using electronic medical records, Fombonne et al. (2021) found 98.8% agreement with SPARK cohort data. SPARK participants are required to have personal access to internet-connected devices to complete studies and surveys online. Written informed consent from and assent was obtained from all participants prior to enrollment.

Of the 105 families that enrolled, 13 did not complete the study, and two had low audio quality such that less than 80% of the adult's speech was intelligible to two trained transcribers. After removing these 15 participants, our sample consisted of 90 parent-child dyads that completed the study between December of 2020 and November of 2021. The 90 child participants (20 female) were between the ages of 4 and 7 years (age in months (M=74.96, SD=12.85, $Range$ = 49–95). **Table 1** shows the racial and ethnic characteristics of the participants and the highest educational degree attained by the child's primary caregiver.

TABLE 1 | Participant demographics.

Characteristic	Number (%)			
Race and Ethnicity	**Not hispanic or latino**	**Hispanic or latino**	**Not reported**	**Total**
American Indian/Alaska Native	1 (1.1%)	0	0	1 (1.1%)
Asian	4 (4.4%)	0	0	4 (4.4%)
Native Hawaiian or Other Pacific Islander	0	0	0	0
Black or African American	5 (5.6%)	1 (1.1%)	0	6 (6.7%)
White	58 (64.4%)	8 (9%)	0	66 (73.3%)
More than one race	5 (5.6%)	2 (2.2%)	1 (1.1%)	8 (9%)
Other	1 (1.1%)	2 (2.2%)	0	3 (3.3%)
Not reported	1 (1.1%)	0	1 (1.1%)	2 (2.2%)
Total	75 (83.3%)	13 (14.5%)	2 (2.2%)	90 (100%)
Primary Caregiver Highest Degree				
High school graduate or GED				6 (6.7%)
Special training after high school (vocational or trade degree)				3 (3.3%)
Some college				18 (20%)
College degree				32 (35.6%)
Graduate or professional degree				29 (32.2%)
No answer				2 (2.2%)
Total				90 (100%)

2.2. Procedure

The Parent-Child Interaction (PCI) consisted of a 15-min, naturalistic interaction between the child and a parent. This interaction was recorded by an examiner over Zoom. Parents were instructed before the interaction to prepare two to four activities that they thought would hold their child's attention for this duration of the interaction and elicit communication. Parents were provided with instructions that included a list of possible activities. Parents were given an opportunity prior to the interaction to brainstorm possible activities with the examiner if they were unsure what would work well with Zoom video. Parents were instructed, if possible, to avoid activities that featured electronic devices and/or toys that made a lot of sounds as these could both discourage active engagement and make existing communication inaudible. They were also instructed to, when possible, interact at a table in a room with minimal distractions and no other people present (see **Supplementary Materials I** for the written instructions that were provided to parents).

Once activities were determined, parents positioned themselves so that both they and their child were visible on screen. The examiner recording the interaction turned off their video so as to not be a distraction, but remained on the call. This allowed the examiner to pause the recording whenever the child needed a break, if there were technical issues with the video call, or to request that the parent and or child reposition

themselves remain visible onscreen. Once 15 min of interaction were recorded, the examiner turned their video back on and informed the caregivers that the interaction was finished.

Parents were provided with detailed step-by-step instructions for downloading and using Zoom, though most families were already familiar with using Zoom. Parents were also provided with detailed instructions for recording high quality audio in wav format using the Lexis audio editor app on a home device and uploading the files to a secure shared folder. Once the parent uploaded the .wav audio file, the research technician moved it to a secure password protected lab server and deleted the file from the shared folder. Recording via the Lexis app on an in-home local device, in addition to Zoom, ensured a second back-up audio file of higher quality, as it did not rely on variable internet connectivity.

We selected a 15-min interaction as previous NLS research has shown that language samples of 10–20 min in length are sufficient to extract reliable language measures from children with autism (Tager-Flusberg et al., 2009; Kover and Abbeduto, 2010). Our piloting showed that a 15-min parent-child interaction was well-tolerated by children with ASD and their parents, who were tasked with keeping the children visible on screen. We chose to prioritize video data, over audio-only, so that we could code video for non-verbal communication, joint attention and engagement for subsequent studies with these data. While manual transcription and coding of these data are labor-intensive, they result in a rich data set. Moreover, reliable automated methods for the analysis of speech in children with ASD over the age of 5 have not yet been developed. A recent test of the reliability of LENA (Language Environment Analysis; Gray et al., 2007), a portable, digital language processor validated for use with the typically developing 0-4 age group, found this method was unreliable for children with autism over the age of 5 (Jones et al., 2019).

Parents also completed the *Vineland Adaptive Behavior Scales-Third Edition* (VABS) (Sparrow et al., 2016) semi-structured interview, an individually administered measure of adaptive functioning used in the diagnosis of intellectual and developmental disabilities. The VABS interview was administered remotely using Zoom by research-reliable technicians. Core domain standard scores represent an examinees overall adaptive functioning across four broad domains: communication, daily living skills, motor skills and socialization. The overall level of adaptive functioning is based on the Adaptive Behavioral Composite (M = 100; SD = 15). Adaptive raw scores were computed at the subdomain level and converted to v-scale scores (M=15; SD=3). **Table 2** shows the children's scores overall and in the four domains assessed.

TABLE 2 | Characteristics of child participants.

Characteristic	*M*	*SD*	Range
Age in months	74.96	12.85	49-95
VABS Standard Score	58.3	13.79	31-84
VABS Communication Domain	53.1	21.49	20-94
VABS Living Domain	63.18	12.76	31-102
VABS Social Domain	57.82	14.12	32-90
VABS Motor Domain	70.14	13.47	20-100

2.3. Transcription

The parent-child interactions were transcribed using the Systematic Analysis for Language Transcripts (SALT; Miller and Iglesias, 2012) procedures. In accordance with SALT procedures, utterances were segmented into communication units defined as an independent clause with its modifiers. A word was defined as a set of characters bound by spaces. Common phrases with co-occurring words that were spoken without pauses between them (e.g., "alldone," "nothankyou," "allgone," "cleanup," "gimme," and "kinda") were transcribed as one word following transcription standards for children with ASD (Tager-Flusberg and Anderson, 1991; Tager-Flusberg et al., 2009; La Valle et al., 2020). Words were transcribed using standard orthography to avoid increasing the number of different words used within and across transcripts. One researcher transcribed all utterances and marked bound morphemes according to SALT conventions. A second researcher then reviewed the file to proof the transcription. Transcription proofing involved reviewing the initial transcript while viewing the video of the parent-child interaction. Discrepancies were settled by the two researchers reaching a consensus in accordance with SALT conventions (Miller and Iglesias, 2012). In the rare case that a consensus could not be reached, the word or utterance in question was marked as unintelligible to avoid inflating the number of intelligible words produced.

All intelligible verbal utterances were included (including utterances that were interrupted or abandoned), since our focus was on number of different words rather than utterances at the conversational level. Unintelligible and nonverbal utterances were excluded. Following conventions for NLS with individuals with ASD (see e.g., Tager-Flusberg and Anderson, 1991; La Valle et al., 2020), we did not include stereotyped language (e.g., echolalia, scripted recitation and idiosyncratic language), sign language or alternative and augmentative communication (AAC) (e.g., speech generating devices). While AAC and manual sign are valid forms of communication, it is unclear how to treat the use of AAC and manual sign, as NLS was developed to analyze spoken language. Similarly, stereotyped language is common in those with ASD, particularly those with lower language levels (La Valle et al., 2020) and serves communicative functions (Stiegler, 2015). Additional studies are needed to understand the role of stereotyped language in language production in children with ASD. Future studies are needed to understand non-spoken communication modalities and stereotyped language use in the context of NLS.

2.3.1. Number of Different Words (NDW)

Number of different words is a well-established measure of lexical diversity (vocabulary development) that can be reliably obtained from a 10 to 15 min language sample for the purpose of screening and/or diagnosis (Miller et al., 2011; Paul et al., 2018). NDW is an optimal measure of language for children with ASD, particularly those who are MLV and have little spoken language (Barokova et al., 2020). We used the SALT

TABLE 3 | Characteristics of child spoken language.

Characteristic	*M*	*SD*	Range
Utterances per minute	2.56	3.09	0–13.27
Percent intelligible utterances	53.74	30.07	0–100
Number of different words	38.81	49.91	0–233
Mean length of utterance in morphemes	1.67	1.24	1–5.94
Time of interaction	14.98	0.13	13.87–15

software (Miller and Iglesias, 2012) to obtain the measure of number of different words. We included only utterances that were complete, intelligible and spontaneous. While stereotyped language is common in children with ASD (Stiegler, 2015), it is typically excluded from NLS measures of spontaneous language ability (see e.g., Tager-Flusberg and Calkins, 1990; Tager-Flusberg and Anderson, 1991). Stereotyped utterances were defined as repetitions, scripted recitations, neologisms and idiosyncratic speech. Repetitions were further defined as that was a complete or partial repetitions of a previous utterance within the past five utterances spoken by either the child or the parent (Tager-Flusberg and Anderson, 1991; La Valle et al., 2020). Singing, reading, counting and other forms of language recitation are typically not considered spontaneous spoken language. While we included those activities in our analyses, we did not include the child's utterances that involved singing, reading, counting or reciting in the spontaneous spoken language measures. In **Table 3**, we outlined a range of spoken language measures for our sample, including talkativeness (number of utterances per minute), speech sound production (percent intelligible utterances), mean length of utterance in morphemes (MLUm) (syntax), and our main measure of vocabulary–number of different word roots (NDW). We also show the time of the interaction because four of the parent-child interactions were under 15 min in length because the child would no longer remain on the video call.

2.4. Coding

2.4.1. Activities

We descriptively categorized the activities in all parent-child interactions based on the materials used and the primary purpose of the activity. Two research assistants categorized all activities for each parent-child dyad. Since the coding of activity categories was primarily descriptive, the activity categorization was then reviewed by the first author, and questions and discrepancies were settled by consensus. The activities were placed into one and only one category based on the following descriptions (with examples):

1. *Conversation only*: No activity or items are presented. The caregiver and child have conversation about themselves or things in their immediate environment.
   Example 1: The child sits on his father's lap at the kitchen table. The child plays with the father's wrist watch, and they talk about it.
   Example 2: The mother asks the child questions about the child's day at school.

2. *Cooking, baking*: Making real food using kitchen items.
   Example 1: The mother gives the boy a cup of whipped cream. The child adds food coloring and stirs. The mother instructs the child to spread the whipped cream on cookies then put sprinkles on them.
   Example 2: The mother and the child make brownies together.

3. *Coloring, art*: Using crayons, markers, paint or other supplies to make a drawing, painting or other craft.
   Example 1: The mother and child draw pictures with markers.
   Example 2: The child colors on her arms and legs with washable markers while the mother comments.

4. *Educational activities*: Activities (including paper/workbooks, flashcards, apps and games) that are explicitly designed to promote literacy or math.
   Example 1: The mother tells the child words to write on a small dry-erase board.
   Example 2: The mother and the child work on math homework sent home by the child's classroom teacher.

5. *Figure play*: Play with action figures, stuffed animals or other toys that can be animated.
   Example 1: The mother and child play with superhero figures making them fly and talk.
   Example 2: The mother and the child play with stuffed animals, putting clothing on them and discussing it.

6. *Games, puzzles*: Turn-taking games and puzzles.
   Example 1: The mother and child put together a puzzle.
   Example 2: The mother and child play the card game Uno.

7. *Manipulatives*: Play with toys that are designed to be manipulated with the hands.
   Example 1: The mother and the child build a tower with blocks.
   Example 2: The father and the child play with lego bricks making enclosures for lego animals

8. *Motor*: Activities that primarily involve gross or fine body movements.
   Example 1: The mother sits on the floor with the child on the couch. The mother reaches for the child's feet and the child pulls them up so the mother can't get them.
   Example 2: The mother and the child throw a ball back and forth.

9. *Screentime*: The child is using a tablet or phone (not used as a communication device or an educational app).
   Example 1: The child is playing with an app on the tablet (not educational or communicative)
   Example 2: The child watches a video on his mother's phone.

10. *Sensory*: Activities that involve the senses, e.g., touch, sight, hearing, taste, smell.

Example 1: The mother sprays different scented spray bottles and the child smells them.
Example 2: The mother and the child play with kinetic sand, forming it into mounds and pushing their fingers into it.

11. *Shared book reading*: The caregiver and child read, look at, comment on, turn the pages of a book together.
Example 1: The mother reads the book and the child comments and turns the pages.
Example 2: The father and the child take turns reading a book together.

12. *Singing, reciting*: Verbal social routines that include songs, counting, reciting the alphabet, reciting poems or riddles.
Example 1: The mother and the child sing *Baby Shark* together.
Example 2: The father helps the child count money the child got for a birthday.

13. *Snack*: Eating a snack is the primary activity.
Example 1: The father gets fruit snacks, giving them to the child one-by-one and prompting the child to ask for more.
Example 2: The mother gives the child fruit snacks one-by-one asking what color the child wants.

2.4.2. Activity Time Range

In order to understand differences between activities, we noted the time range that each parent-child dyad spent engaged in a particular activity. Two research assistants annotated the start time and end time of the activity based on when the parent presented the activity materials (start time) and when the materials were put away or put aside (end time). If a child rejected the activity, it was not counted. Only activity durations longer than 20 s were included, as many activities shorter than 20 s did not engage the child, an alternative activity was presented.

2.4.3. Number of Different Words (NDW per Minute by Activity Type)

Transcriptions were marked with the start and end time of the activity. Using SALT, we extracted NDW for the duration of the activity by specifying the start time and end time in the SALT settings. Then, we calculated NDW per minute by dividing NDW by the duration of the activity to standardize the measure across activities with varying durations.

3. ANALYSIS

3.1. What Activities Do Parents Choose to Promote Spoken Language With Their Children in a Remote Context?

Our first aim was to understand the range of activities that parents selected to promote communication in the home with their child with ASD. **Table 4** shows the percentage of parent-child dyads that engaged in each type of activity. The most frequent activities were: sensory activities, play with manipulative toys, conversation only, games or puzzles, coloring or other art activities, snack, play with toy figures and shared book reading. Less frequent activities include educational (math or literacy) activities, motor activities, screentime, singing or reciting and cooking or baking.

TABLE 4 | Percentage of parent-child dyads engaged in different activity types.

Activity type	Number	Percentage
Manipulatives	27	30
Games and puzzles	23	25.6
Sensory	21	23.3
Shared book reading	18	20
Coloring and art	16	17.8
Figure play	16	17.8
Conversation only	15	16.7
Motor	14	15.6
Educational	10	11.1
Snack	10	11.1
Singing and reciting	9	10.4
Screen time	6	6.7
Cooking and baking	3	3.3

3.2. Does the Type of Activity or Number of Activities Depend on Child Language Level?

Figure 1 shows the child's NDW by activity type. We conducted a Chi-square test to evaluate if there was a significant difference in the mean NDW between the different activity types, however the difference was not significant [$\chi^2(540) = 484.48, p = 0.96$]. Children with lower language, those who had fewer than 20 different spoken words during the parent-child interaction, engaged in all types of activities. While there was not a significant difference across activities, **Figure 1** shows that dyads with children whose NDW levels were higher than 80 different words in 15 min did not engage in coloring and art, cooking and baking, motor activities, screentime, singing and reciting or snack activities.

Figure 2 shows the child's NDW and the number of activities in which the parent-child dyad engaged. A simple linear regression model showed that the child's NDW was a significant predictor of the number of activities ($\beta = -0.01, t = -3.52, p < 0.001$). Parent-child dyads whose children had higher language ability tended to engage in a smaller number of activity types, while parent-child dyads whose children had lower language ability engaged in a range of one to five different activity types.

3.3. Does Type of Activity Elicit More Vocabulary From Children?

Our third aim was to examine if different types of activities elicit more language from children. We found no significant difference in the number of different words per minute elicited during the five most common activities: coloring/art, games/puzzles, play with manipulatives, sensory activities and shared book reading [$\chi^2(264) = 270.55, p = 0.38$] (see **Figure 3**).

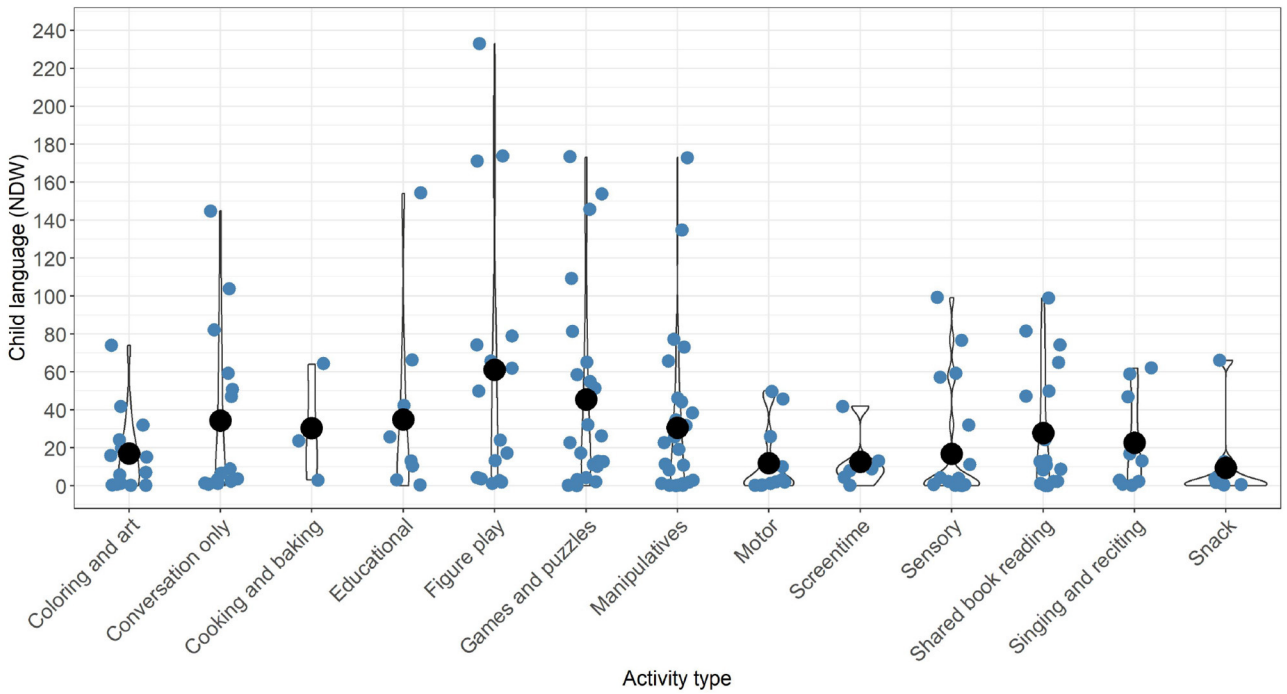

FIGURE 1 | Activity types and child spoken language level.

4. DISCUSSION

4.1. Range of Parent-Selected Activities

The focus of this study was on remote assessment of child language using open ended-parent-child interactions in the home. Parents were not provided with a set of specific activities and materials but were allowed to use the materials in the home and the activities that their children preferred. We found that parents chose activities that fell into 13 different descriptive categories, including some activities that are not typically included in lab-based semi-structured language assessments, such as cooking and baking. We found that all categories of activity were used with children who had low language levels. For dyads with children whose language level was higher (e.g., NDW above 80), they did not engage in coloring and art, cooking and baking, motor activities, screentime, singing and reciting or snack activities. They did engage in conversation only, educational activities, figure play, games and puzzles, manipulatives, sensory activities and shared book reading. Understanding the range of activities that parents select to engage children with ASD in the home with in-home materials and activities is essential to increasing the accessibility and equity of language assessment during pandemic stay-at-home times and beyond.

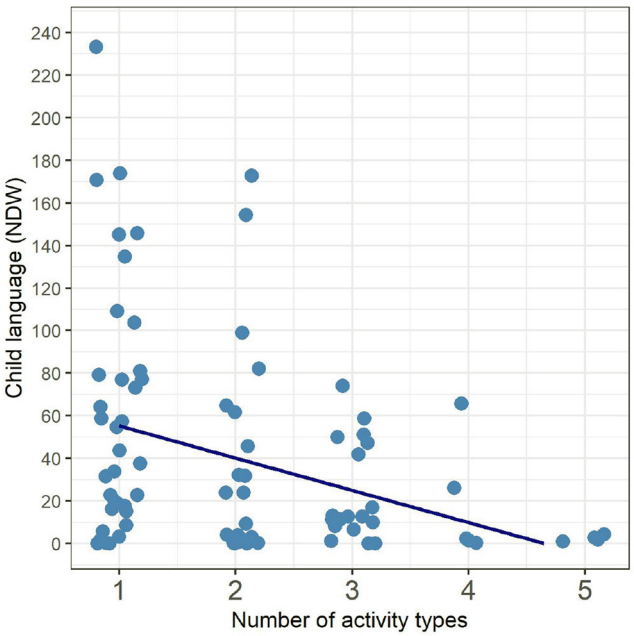

FIGURE 2 | Number of different activity types and child spoken language level.

4.2. Type and Number of Activities

We then aimed to discover the relationship between the child's language level and the type and number of activities. Not only are parents of lower verbal children with ASD engaging in all types of activities, but we also found that there was no significant difference in child language level across activities. Parents of nonverbal and minimally verbal children did not engage in different activities from those whose children had

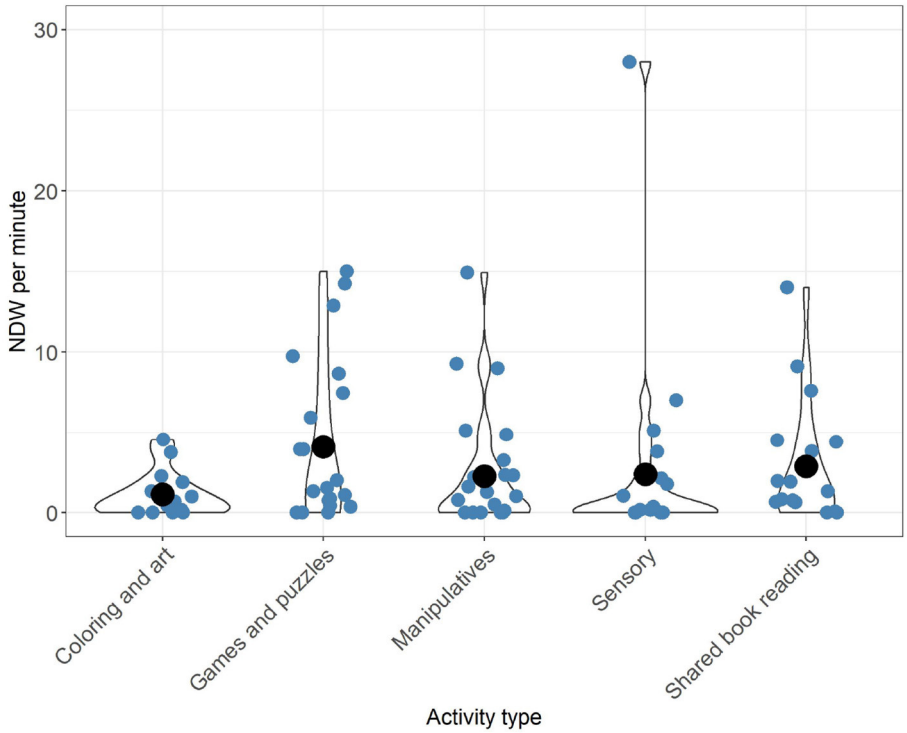

FIGURE 3 | Number of different words (NDW) per minute by activity type.

higher language levels. We did find a significant relationship between the child's language level and the number of different activity types. Parents of children with low language levels tended to engage in a greater number of different activities (up to five in the 15-min interaction). In a recent study using NLS for language assessment with children with ASD, Barokova and colleagues (Barokova et al., 2020) developed a novel protocol for eliciting natural language samples with minimally and low verbal children with ASD. Their protocol involved eight different activities designed to be elicited in a 20-min timeframe. This approach aligns with our findings that parent-child dyads with children who had lower language levels engaged in a higher number of different activity types.

Some, but not all, parent-child dyads in which the child had a smaller number of different words, engaged in a larger number of different types of activities. It is well-known that children with autism have significant social-communication delays in symbolic play and joint attention that differentiate them from typically developing children and children with intellectual disability without autism (Mundy et al., 1986). Both symbolic play and joint attention are significantly associated with social (Sigman et al., 1999), cognitive (Mundy et al., 2010) and communication development (Kasari et al., 2008) as well as expressive langauge in particular (Adamson et al., 2019). Therefore, the level of the child's skills in symbolic play and joint attention may have played a role mediating the relationship between the number of different activities and the language level of the child.

Other factors may have contributed to the association between number of activities and number of different words. The environmental setup of certain activities may not have been feasible within the camera frame for some families. The families may not have had access to all the necessary activity materials during the PCI, since they were asked to do the activity at a tabletop in a quiet room. Finally, while some parents may have added different types of activities to keep their children with lower language engaged, we observed a broader range of engagement strategies that parents used. Along similar lines, parents are in tune to their child's play skills and abilities, which plays a role in the child's response to the interaction (Barokova et al., 2020). These factors should be considered in future research on approaches to analyzing naturalistic parent-child interactions.

4.3. Do Different Activity Types Elicit More Language?

Our third question was whether different activities elicit more language (measured by number of different words per minute) from children with ASD. We found that there was not a significant difference in NDW per minute in the five more common activities: coloring/art, games/puzzles, play with manipulatives, sensory activities and shared book reading. Similar to our findings that type of activity did not elicit significantly more language, Barokova et al. (2020) found no significant difference in spoken language production (measured by frequency of utterances per minute) between activities (with the exception of watching a short animated movie designed to elicit a narrative or naming of the movie characters). Both our study and the Barokova study reported no difference in spoken language production between activities in their protocol, with the

possible exception of screentime. Taken together, these results suggest that a wide range of items, materials and activities, including those already in the home for remote NLS, do not significantly affect the quality or quantity of language elicited from the child.

Rather than play activities being determined by the child's language level, it is possible that play activities are more highly influenced by the child's level of symbolic play. As previously discussed, delays in symbolic play in children with autism are associated with social, cognitive and communication development (Sigman et al., 1999; Kasari et al., 2008; Mundy et al., 2010). Symbolic play allows children to progress developmentally from playing with toys functionally, such as in constructive and manipulative play, to playing with toys symbolically, such as in figurative play (Lifter et al., 1993). Compared to typically developing children matched on mental age, children with autism have significant delays in the development of symbolic play (Baron-Cohen, 1987; Jarrold et al., 1993). Children with autism show less spontaneous, creative symbolic play (Jarrold et al., 1993; Libby et al., 1998) and more manipulation of objects in a rigid or stereotyped manner (Atlas, 1990). Beyond these delays in play skills, children with autism show more focus on objects with less frequent engagement of others into their play activities (Kasari et al., 2010). It is likely that symbolic play skills in children is more predictive of choice of activity than language level, and future work should examine the role of joint attention and symbolic play in remote, open ended-parent-child interactions in the home.

5. LIMITATIONS AND FUTURE DIRECTIONS

Following NLS conventions for individuals with ASD (see e.g., Tager-Flusberg and Anderson, 1991; La Valle et al., 2020), our analyses did not include stereotyped language (e.g., echolalia, scripted recitation, and idiosyncratic language), sign language or alternative and augmentative communication (AAC) (e.g., speech generating devices). While AAC and sign language are valid forms of communication, and some children in the sample appeared to use AAC spontaneously, it is unclear how to treat the use of AAC and sign, as NLS was developed to analyze spoken language. Similarly, stereotyped language is common in those with ASD, particularly those with lower language levels (La Valle et al., 2020) and serves communicative functions (see e.g., Stiegler, 2015). Future studies are needed to understand the use of AAC, sign language and stereotyped language. It is important to understand how to analyze use of AAC, sign language and stereotyped language using NLS and to examine how these influence the development of language in children with ASD.

Another limitation and potential future direction involves the categorization of activity types. Effects may have been different if activity types were grouped differently. For example, there are clear similarities between some activities categorized as sensory, motor, manipulatives and figure play. While playing with blocks and legos was categorized as manipulatives, playing with playdoh or kinetic sand was considered sensory play. However, if the parent-child dyad was playing with playdoh and figures and the primary purpose of the play involved interactions between the figures, then the activity was categorized as figure play. Similarly, for a child whose parent gave him or her small marshmallow rings to string on a straw, this activity was categorized as motor due to the fine motor focus of the activity, but it could have been considered a manipulative activity. In addition, some activity categories overlapped, such as the previous example of play that involved playdoh and toy figures. We restricted our activity coding to a single activity, but a deeper understanding of simultaneous activities would improve our understanding of naturalistic parent-child interactions for the purpose of remote language assessment. Along similar lines, understanding the complexity of play skills (see e.g., Bornstein and O'Reilly, 1993; Freeman and Kasari, 2013; Bentenuto et al., 2016) and parent strategies for responding to and engaging their child (see e.g., Adamson et al., 2012, 2019) was beyond the scope of this paper, but will improve our understanding of methods for remote naturalistic language sampling for children with ASD.

6. CONCLUSIONS

In this study, we analyzed natural language samples of naturalistic interactions between parents and children with ASD collected remotely in the home. We gave parents few parameters to allow for naturalistic play-based interactions in the home with items, materials and activities in the family's own home. It is important to understand naturalistic, open-ended remote NLS because they open the window to accessible and equitable methods for language assessment, particularly for children with ASD who are nonverbal and minimally verbal, for whom current standardized language assessments are not feasible or valid. While parent-child dyads engaged in a wide range of different activity types with their children in the home, we found no significant difference between activity type and the child's language level measured by NDW. Parents of children who were nonverbal and minimally verbal engaged in all types of activities, and they engaged in the same activities as did parents of children with higher language levels. We did find, however, a significant relationship between language level and the number of different activity types. Parents of children with lower language levels tended to engage in a higher number of different activity types. Different activities did not elicit significantly more language. These results suggest that remote, in-home NLS with items, materials and activities selected by parents are an appropriate method to assess language remotely in children with ASD, so long as a sufficient number of activity types are presented to children who have lower language levels. Finally, our results support the feasibility of remote in-home natural language sampling using the family's own items, materials and activities.

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: National Database for Autism Research (NDAR).

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

LB contributions include transcription, coding, conceptualization, analysis, writing, and editing. CLa contributed conceptualization, transcription, and editing. SS contributed conceptualization, data acquisition, and editing. JP contributed data acquisition and writing. CLi contributed transcription and coding. NP and LS contributed transcription. HT-F contributed funding acquisition, lab resources, conceptualization, and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Institutes of Health P50DC018006 (PIs Tager-Flusberg/Kasari). REDCap electronic data capture tools hosted at Boston University were established with a grant to the Clinical and Translational Science Institute (1UL1TR001430).

ACKNOWLEDGMENTS

We thank the Simons Foundation Powering Autism Research for Knowledge (SPARK) research match team for assistance designing the protocol to recruit participants for this study. We thank the families who gave their time to participate in this research. We thank members of the Center for Autism Research Excellence for their comments on earlier version of this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.820564/full#supplementary-material

REFERENCES

Adamson, L. B., Bakeman, R., Deckner, D. F., and Nelson, P. B. (2012). Rating parent-child interactions: Joint engagement, communication dynamics, and shared topics in autism, Down syndrome, and typical development. *J. Autism. Dev. Disord*. 42, 2622–2635. doi: 10.1007/s10803-012-1520-1

Adamson, L. B., Bakeman, R., Suma, K., and Robins, D. L. (2019). An expanded view of joint attention: Skill, engagement, and language in typical development and autism. *Child Dev*. 90, e1–e18. doi: 10.1111/cdev.12973

American Communities Survey (2019). *Types of computers and internet subscriptions. Technical report, U. S. Census Bureau*. Available online at: https://data.census.gov/cedsci/table?q=smartphoneandtid=ACSST1Y2019.S2801.

Atlas, J. A. (1990). Play in assessment and intervention in childhood psychoses. *Child Psychiatry Hum. Dev*. 21, 119–133. doi: 10.1007/BF00706120

Barokova, M., La Valle, C., Hassan, S., Lee, C., Xu, M., McKechnie, R., et al. (2020). Eliciting language samples for analysis (ELSA): a new protocolfor assessing expressive language and communication in autism. *Autism Res*. 14, 112–126. doi: 10.1002/aur.2380

Barokova, M. D., and Tager-Flusberg, H. (2018). Commentary: measuring language change through natural language samples. *J. Autism Dev. Disord*. 50, 2287–2306. doi: 10.1007/s10803-018-3628-4

Baron-Cohen, S. (1987). Autism and symbolic play. *Br. J. Dev. Psychol*. 5, 139–148. doi: 10.1111/j.2044-835X.1987.tb01049.x

Bentenuto, A., De Falco, S., and Venuti, P. (2016). Mother-child play: a comparison of autism spectrum disorder, down syndrome, and typical development. *Front. Psychol*. 7, 1829. doi: 10.3389/fpsyg.2016.01829

Bornstein, M. H., and O'Reilly, A. W. (1993). *The Role of Play in the Development of Thought*. San Francisco, CA: Jossey-Bass.

Burgess, S., Audet, L., and Harjusola-Webb, S. (2013). Quantitative and qualitative characteristics of the school and home language environments of preschool-aged children with asd. *J. Commun. Disord*.,46, 428–439. doi: 10.1016/j.jcomdis.2013.09.003

Costanza-Smith, A. (2010). The clinical utility of language samples. *Perspect. Lang. Learn. Educ*. 17, 9–15. doi: 10.1044/lle17.1.9

Dash, S., Aarthy, R., and Mohan, V. (2021). Telemedicine during COVID-19 in India–a new policy and its challenges. *J. Public Health Policy* 42, 501–509. doi: 10.1057/s41271-021-00287-w

Evans, J. L., and Craig, H. K. (1992). Language sample collectionand analysis: interview compared to freeplay assessment contexts. *J. Speech Lang. Hear. Res*. 35, 345–353. doi: 10.1044/jshr.3502.343

Feliciano, P., Daniels, A. M., Green Snyder, L., Beaumont, A., Camba, A., Esler, A., et al. (2018). SPARK: a US cohort of 50,000 families to accelerate autism research. *Neuron* 97, 488–493. doi: 10.1016/j.neuron.2018.01.015

Fombonne, E., Coppola, L., Mastel, S., and O'Roak, B. J. (2021). Validation of autism diagnosis and clinical data in the SPARK cohort. *J. Autism Dev. Disord*. doi: 10.1007/s10803-021-05218-y (accessed July 30, 2021).

Franz, L., Chambers, N., von Isenburg, M., and de Vries, P. J. (2017). Autism spectrum disorder in sub-saharan Africa: a comprehensive scoping review. *Autism Res*. 10, 723–749. doi: 10.1002/aur.1766

Freeman, S., and Kasari, C. (2013). Parent-child interactions in autism: characteristics of play. *Autism* 17, 147–161. doi: 10.1177/1362361312469269

Gladfelter, A., and Van Zuiden, C. (2020). The influence of language context on repetitive speech use in children with autism spectrum disorder. *Am. J. Speech Lang. Pathol*. 29, 327–334. doi: 10.1044/2019_AJSLP-19-00003

Gray, S. S., Baer, C. T., Xu, D., and Yapanel, U. (2007). The LENA Language Environment Analysis System: *The Infoture Time Segment (ITS) File*. LENA Foundation, Boulder, CO, United States. Available online at: https://www.lena.org/wp-content/uploads/2016/07/LTR-04-2_ITS_File.pdf

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., McLeod, L., Delacqua, G., Delacqua, F., Kirby, J., and Duda, S. N. (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

Hilvert, E., Sterling, A., Haebig, E., and Friedman, L. (2020). Expressive language abilities of boys with idiopathic autism spectrum disorder and boys with fragile x syndrome + autism spectrum disorder: cross-context comparisons. *Autism Dev. Lang. Impairments* 5, 1–16. doi: 10.1177/2396941520912118

Jarrold, C., Boucher, J., and Smith, P. K. (1993). Symbolic play in autism: a review. *J. Autism Dev. Disord*. 23, 281–307. doi: 10.1007/BF01046221

Jones, R. M., Plesa Skwerer, D., Pawar, R., Hamo, A., Carberry, C., Ajodan, E. L., et al. (2019). How effective is LENA in detecting speech vocalizations and language produced by children and adolescents with ASD in different contexts? *Autism Res*. 12, 628–635. doi: 10.1002/aur.2071

Kasari, C., Gulsrud, A. C., Wong, C., Kwon, S., and Locke, J. (2010). Randomized controlled caregiver mediated joint engagement intervention for toddlers with autism. *J. Autism Dev. Disord.* 40, 1045–1056. doi: 10.1007/s10803-010-0955-5

Kasari, C., Paparella, T., Freeman, S., and Jahromi, L. (2008). Language outcomes in autism: randomized comparison of joint attention and play interventions. *J. Consult Clin. Psychol.* 76, 125–137. doi: 10.1037/0022-006X.76.1.125

Kover, S. T., and Abbeduto, L. (2010). Expressive language in male adolescents with fragile X syndrome with and without comorbid autism. *J. Intell. Disabil. Res.* 54, 246–265. doi: 10.1111/j.1365-2788.2010.01255.x

Kover, S. T., Davidson, M. M., Sindberg, H. A., and Weismer, S. E. (2014). Use of the ADOS for assessing spontaneous expressive language in young children with asd: a comparison of sampling contexts. *J. Speech Lang. Hear. Res.* 57, 2221–2233. doi: 10.1044/2014_JSLHR-L-13-0330

La Valle, C., Plesa-Skwerer, D., and Tager-Flusberg, H. (2020). Comparing the pragmatic speech profiles of minimally verbaland verbally fluent individuals with autism spectrum disorder. *J. Autism Dev. Disord.* 50, 3699–3713. doi: 10.1007/s10803-020-04421-7

Libby, S. S., P., Messer, D., and Jordan, R. (1998). Spontaneous play in children with autism: a reappraisal. *J. Autism Dev. Disord.* 28, 487–497. doi: 10.1023/A:1026095910558

Lifter, K., Sulzer-Azarnoff, B., Anderson, S. R., and Cowdery, G. E. (1993). Teaching play activities to preschool children with disabilities: the importance of developmental considerations. *J. Early Interv.* 17, 139–159. doi: 10.1177/105381519301700206

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., and Bishop, S. L. (2012). *Autism Diagnostic Observation Schedule, 2nd Edn.* Torrance, CA: Western Psychological Services.

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. Hearing Research*, 62, 3982–3990. doi: 10.1044/2020_JSLHR-20-00202

Miller, J. (1981). *Assessing language production in children: Experimental procedures*. Allyn and Bacon, Boston, MA.

Miller, J., Andriacchi, K., and Nockerts, A. (2011). *Assessing Language Production Using SALT Software: A Clinician's Guide to Language Sample Analysis*. Madison, WI: SALT Software, LLC.

Miller, J., and Iglesias, A. (2012). *Systematic analysis of language transcripts (SALT). Computer software, SALT Software*. Research Version 2012.

Mundy, P., Gwaltney, M., and Henderson, H. (2010). Self-referenced processing, neurodevelopment and joint attention in autism. *Autism* 14, 408–429. doi: 10.1177/1362361310366315

Mundy, P., Sigman, M., Ungerer, J., and Sherman, T. (1986). Defining the social deficits of autism: the contribution of non-verbal communication measures. *J. Child Psychol. Psychiatry* 27, 657–669. doi: 10.1111/j.1469-7610.1986.tb00190.x

Paul, R., Norbury, C., and Gosse, C. (2018). *Language Disorders From Infancy Through Adolescence, 5th Edn*. St. Louis, MO: Elsevier.

Rosen, N., Lord, C., and Volkmar, F. (2021). The diagnosis of autism: From Kanner to DSM-III to DSM-5 and beyond. *J. Autism Dev. Disord.* 51, 4253–4270. doi: 10.1007/s10803-021-04904-1

Sanchez, K., Spittle, A. J., Boyce, J. O., Leembruggen, L., Mantelos, A., Mills, S., et al. (2020). Conversational language in 3-year-old children born very preterm and at term. *J. Speech Lang. Hear. Res.* 63, 206–215. doi: 10.1044/2019_JSLHR-19-00153

Schreibman, L., Dawson, G., Stahmer, A. C., Landa, R., Rogers, S. J., McGee, G. G., et al. (2015). Naturalistic developmental behavioral interventions: empirically validated treatments for autism spectrum disorder. *J. Autism Dev. Disord.* 45, 2411–2428. doi: 10.1007/s10803-015-2407-8

Sigman, M., Ruskin, E., Arbeile, S., Corona, R., Dissanayake, C., Espinosa, M., Kim, N., López, A., and Zierhut, C. (1999). Continuity and change in the social competence of children with autism, down syndrome, and developmental delays. *Monogr. Soc. Res. Child Dev.* 64, 1–114. doi: 10.1111/1540-5834.00010

Sparrow, S. S., Cicchetti, D. V., and Saulnier, C. A. (2016). *Vineland Adaptive Behavior Scales, 3rd Edn*. San Antonio, TX: Pearson.

Stiegler, L. N. (2015). Examining the echolalia literature: where do speech-language pathologists stand? *Am. J. Speech Lang. Pathology* 24, 750–762. doi: 10.1044/2015_AJSLP-14-0166

Tager-Flusberg, H., and Anderson, M. (1991). The development of contingent discourse ability in autistic children. *J. Child Psychol. Psychiatry* 32, 1123–1134. doi: 10.1111/j.1469-7610.1991.tb00353.x

Tager-Flusberg, H., and Calkins, S. (1990). Does imitation facilitate the acquisition of grammar? evidence from a study of autistic, down syndrome and normal children. *J. Child Lang.* 17, 591–606. doi: 10.1017/S0305000900010898

Tager-Flusberg, H., and Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorders: The neglected end of the spectrum. *Autism Res.* 6, 468–478. doi: 10.1002/aur.1329

Tager-Flusberg, H., Rogers, S., Cooper, J., Landa, R., Lord, C., Paul, R., et al. (2009). Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. *J. Speech Lang. Hear. Res.* 52, 643–656. doi: 10.1044/1092-4388(2009/08-0136)

Zwaigenbaum, L., Bishop, S., Stone, W. L., Ibanez, L., Halladay, A., Goldman, S., et al. (2021). Rethinking autism spectrum disorder assessment for children during COVID-19 and beyond. *Autism Res.* 14, 2251–2259. doi: 10.1002/aur.2615

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.

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 22 July 2022
DOI 10.3389/fcomm.2022.798196

OPEN ACCESS

EDITED BY
Wenchun Yang,
Leibniz Center for General Linguistics (ZAS), Germany

REVIEWED BY
Eliseo Diez-Itza,
University of Oviedo, Spain
Jiangling Zhou,
The Chinese University of Hong Kong, China

*CORRESPONDENCE
Diana Burchell
diana.burchell@mail.utoronto.ca

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 19 October 2021
ACCEPTED 04 July 2022
PUBLISHED 22 July 2022

CITATION
Burchell D, Bourassa Bédard V, Boyce K, McLaren J, Brandeker M, Squires B, Kay-Raining Bird E, MacLeod A, Rezzonico S, Chen X, Cleave P and FrEnDS-CAN (2022) Exploring the validity and reliability of online assessment for conversational, narrative, and expository discourse measures in school-aged children. *Front. Commun.* 7:798196. doi: 10.3389/fcomm.2022.798196

Exploring the validity and reliability of online assessment for conversational, narrative, and expository discourse measures in school-aged children

Diana Burchell[1]*, Vincent Bourassa Bédard[2], Keara Boyce[3], Juliana McLaren[3], Myrto Brandeker[3], Bonita Squires[3], Elizabeth Kay-Raining Bird[3], Andrea MacLeod[4], Stefano Rezzonico[2], Xi Chen[1], Pat Cleave[3] and FrEnDS-CAN

[1]Ontario Institute for Studies in Education, University of Toronto, Toronto, ON, Canada, [2]École d'orthophonie et d'audiologie, University of Montréal, Montreal, QC, Canada, [3]School of Communication Sciences and Disorders, Dalhousie University, Halifax, NS, Canada, [4]Faculty of Rehabilitation Medicine, University of Alberta, Edmonton, AB, Canada

The COVID-19 pandemic has created novel challenges in the assessment of children's speech and language. Collecting valid data is crucial for researchers and clinicians, yet the evidence on how data collection procedures can validly be adapted to an online format is sparse. The urgent need for online assessments has highlighted possible the barriers such as testing reliability and validity that clinicians face during implementation. The present study describes the adapted procedures for on-line assessments and compares the outcomes for monolingual and bilingual children of online and in-person testing using conversational, narrative and expository discourse samples and a standardized vocabulary test. A sample of 127 (103 in-person, 24 online) English monolinguals and 78 (53 in-person, 25 online) simultaneous French-English bilinguals aged 7–12 years were studied. Discourse samples were analyzed for productivity, proficiency, and syntactic complexity. MANOVAs were used to compare on-line and in-person testing contexts and age in two monolingual and bilingual school-age children. No differences across testing contexts were found for receptive vocabulary or narrative discourse. However, some modality differences were found for conversational and expository. The results from the study contribute to understanding how clinical assessment can be adapted for online format in school-aged children.

KEYWORDS

discourse measures, speech-language pathology, online assessment, children, conversation, expository, narrative

Introduction

During the COVID-19 pandemic, practitioners have been searching for feasible adaptations to current language assessment practices, since traditional in-person assessments have not been possible. Being able to collect valid data is crucial for researchers and clinicians, yet the evidence on how data collection procedures specific to language samples can validly be adapted to an online format is sparse (Taylor et al., 2014; d'Orville, 2020; Reimers et al., 2020). Online assessments of language, traditionally used for remote clients, have now been widely implemented by researchers and clinicians alike due to the sudden lockdown of in-person services in the beginning of 2020 (e.g., Mansuri et al., 2021). Emerging evidence is looking at the developments necessary to make tele-practice a reliable and valid assessment method (Chenneville and Schwartz-Mette, 2020; Putri et al., 2020; Fong et al., 2021). The present study will describe the adapted procedures required to pursue a large-scale study of discourse in typically developing monolingual and bilingual school-aged children. We will examine whether online and in-person conversation, narrative or expository discourse sample measures or standardized vocabulary tests differ for monolingual and bilingual children, and, if so, what might account for these differences.

Few studies have investigated the online data collection of discourse samples among monolingual and bilingual school-aged children. There are a few studies on monolingual adults in clinical populations (e.g., Turkstra et al., 2012), as well as studies of both monolingual (Manzanares and Kan, 2014) and bilingual (Guiberson et al., 2015) preschool children. One systematic review by Taylor et al. (2014) evaluated the efficacy and effectiveness of speech and language assessments online. The authors found 5 studies who met the inclusion criteria but stated that the articles were of variable quality and did not provide enough evidence to influence clinical practice. This review presented evidence of inter-rater reliability in online language assessments, and most studies found sufficient inter-rater reliability to suggest the contexts did not significantly alter the results. They confirmed that more rigorous statistics are needed to either confirm or dispute this preliminary evidence. More recently, Manning et al. (2020) looked at the feasibility, reliability, and validity of obtaining language samples remotely by recording child-parent play with toddlers. They compared online and in-person groups on language sample metrics such as mean length of utterance (MLU), number of different words (NDW), and type-token ratio (TTR). This study found no evidence of differences across any language metrics due to modality. To the best of our knowledge, there are few published studies that have investigated the comparability of online and in-person assessments of school-age children using rigorous and parametric statistics. The following paragraphs will provide an overview of research relating to measures of vocabulary, and more in-depth review of measures of discourse and language metrics. While the larger project included both macrostructural (i.e., related to the meaning conveyed) and microstructural (i.e., related to the language used) measures, for the purposes of this paper, only microstructural measures were included. This was a strategic decision since microstructural measures represent language development, which is of particular interest to clinicians during the pandemic.

Vocabulary measures have been previously validated for use in online assessment. Haaf et al. (1999) created and evaluated two online measures of the Peabody Picture Vocabulary Test-Revised (PPVT-R; Dunn and Dunn, 1981). They found that the online versions of the PPVT-R were not significantly different from the in-person version of the PPVT-R. The authors concluded that an online version of the PPVT is statistically equivalent to the in-person version and can be used in conjunction with the published norms. Eriks-Brophy et al. (2008) later confirmed this statistical equivalence using an updated version of the PPVT (PPVT-III, Dunn and Dunn, 1997). This replication of Haaf's original study with a new version of the PPVT indicates the stability of these findings, even with slight methodological changes.

In general terms, discourse is typically defined as a conversation between people as a form of communication. Within the fields of linguistics and speech language pathology, discourse is more specifically defined as 'a linguistic unit (such as conversation or a story) larger than a sentence' [Merriam-Webster, (n.d)]. Discourse skills have been shown to be critical to school success and are known to be an area of difficulty for children with language-learning disabilities (Paul and Norbury, 2012). Three main types of discourse include conversation, exposition, and narration. Conversation has been defined as a "dialogue between people where each contributes by making statements, asking questions, and responding to the other speaker" (Nippold et al., 2014, p. 877). Exposition and narration are monologic in clinical settings, where expository discourse is defined as the use of language to convey information (Bliss, 2002) while narrative discourse is defined as telling stories about oneself and/or others (Nippold et al., 2014).

Children develop discourse skills over a long period of time and across a variety of genres. They begin to develop the ability to engage in conversational discourse even before they start to speak, and this skill continues to be refined through the school years (Hoff, 2009). Expository discourse begins to develop later than conversation and narration. Procedural description, persuasion, negotiation and explanation are all forms of expository discourse (Nippold et al., 2007; Nippold and Sun, 2010). Expository discourse emerges within conversations in the preschool period (Cabell et al., 2011) but becomes more prevalent in children's experiences once schooling begins and is increasingly more frequent in their spoken language at that time (Nippold and Sun, 2010). In narration, children begin to

talk about past events and to produce brief narrative recounts of these events by the age of two when scaffolded by a parent (Eisenberg, 1985). By 5 years of age, they are able to produce narratives with some plot structure (Hoff, 2009; Owens, 2012) and the complexity of their spoken narratives continues to develop through at least 12 years of age (Hoff, 2009; Cabell et al., 2011). The following paragraphs will provide an overview of the literature investigating and comparing language use in these three discourse genres.

Conversational tasks have been the most effective in accurately portraying the discourse level skills of younger children (Leadholm and Miller, 1992; Heilmann et al., 2010). Furthermore, conversational tasks are more reflective of basic interpersonal communication skills (e.g., BICS) as opposed to later developing discourse tasks that may be more in line with curriculum expectations and cognitive academic language proficiency (e.g., CALP), such as expository and narrative measures (Heilmann et al., 2010). However, assessing basic interpersonal language in school-age children may still be useful for clinicians, especially in the case of language learners who may not have developed adequate academic language yet (see Cummins, 2000 for a review of BICS and CALP).

Expository tasks are highly structured measures which focus on explaining a specific topic (Berman and Nir-Sagiv, 2007). Furthermore, expository tasks can be curriculum based, which can be a powerful diagnostic tool in evaluating children's expressive language (Heilmann and Malone, 2014). There is preliminary evidence that expository tasks accurately capture the development of academic language skills (Kay-Raining Bird et al., 2016). This is supported by a study conducted by Nippold et al. (2005), which compared conversational and expository discourse. The authors found that students demonstrated greater syntactic complexity on the expository task, indicating that complex thought underlies complex language. In the present study, looking at an expository measure is useful as an index of both academic language and curriculum-based assessment.

Researchers such as Stadler and Ward (2005) have shown that narrative skills are a rich reflection of children's oral language development. This is because narratives require more complex vocabulary and an overarching structure. Narrative skills are typically assessed in two ways, wherein students are asked to demonstrate comprehension of a "model" story and/or produce an original story. Storytelling skills emerge in the preschool period and continue to grow throughout their time in school. Narrative development is supported through activities such as storybook reading. As students' oral language competency grows, it increases their complexity of their language (Verhoeven and Strömqvist, 2001). In this study, on-line and in-person narrative production tasks were compared.

Discourse tasks provide a context for observing children's language abilities, and specific language metrics allow researchers and clinicians to obtain quantitative observations that can reflect children's development and proficiency. For the purposes of this paper, we chose to include the most commonly used language metrics, which includes measures of productivity, proficiency, and syntactic complexity (Schneider et al., 2004). Lexical productivity is a metric which measures the amount of output generated by a participant (Le Normand et al., 2008). Studies have shown that lexical productivity tends to increase with age for both monolingual and bilingual children (Le Normand et al., 2008; Jia et al., 2014). Productivity was of particular interest in this study since it has been correlated with psycho-social variables such as introversion, anxiety and shyness (see Dewaele and Pavlenko, 2003 for a review). Emerging research during the COVID-19 pandemic has shown increased levels of anxiety and shyness due to online schooling and social isolation (Imran et al., 2020; Lavigne-Cerván et al., 2021; Orgilés et al., 2021). It is therefore vital to ascertain whether psycho-social factors such as anxiety and shyness may affect the administration of online discourse-level skills for clinicians.

Multiple studies attest to the critical role of syntactic complexity in the development of language and literacy skills in school-age populations. However, syntactic complexity is dependent on the type of task administered: studies show that children produce more complex utterances during expository discourse than they do in conversation (e.g., Nippold, 2009). There is also emerging evidence that modeling (used in our narrative assessments), which involves syntactic priming, may also impact syntactic complexity (Zebib et al., 2020). Syntactic complexity is therefore an interesting metric in this study, since it can be variable across tasks but may be stable across modality contexts.

Mean length of C-Unit in morphemes (MLCUm) is a micro-structural (i.e., linguistic) measure that is a general reflection of both general language proficiency as well as the syntactic complexity of the discourse being analyzed (Craig et al., 1998; Eisenberg et al., 2001). An utterance is defined as one main clause and all dependent clauses associated with it (Miller et al., 2006). With increased language proficiency, children will begin to incorporate more advanced linguistic devices into their speech, including conjunctions and subordinate clauses, resulting in greater MLCUm values (Berman and Slobin, 1994). However, in the past, researchers have found that MLCUm changes with age, improving more significantly in expository and narrative discourse than in conversation for older students (i.e., teen years) (Leadholm and Miller, 1992; Rice et al., 2010; Westerveld and Moran, 2013). In this study, MLCUm is one metric used to measure whether children are showing the same proficiency on-line as they would in person. Similar to syntactic complexity, we expect that MLCUm may be stable across modality contexts.

The current study emerged from the project "*French/English Discourse Study – Canada*" (FrEnDS-CAN), which focuses on a variety of discourse skills in typically developing

TABLE 1 Frequency of gender and age group by condition and language group.

		Monolingual		Bilingual	
		In-person ($n = 103$)	Online ($n = 24$)	In-person ($n = 53$)	Online ($n = 25$)
Gender	Male	34	10	22	2
	Female	44	13	19	5
	Not disclosed	25	1	12	18
Age group	7–8 years	48	5	19	10
	9–10 years	38	10	18	10
	11–12 years	17	9	16	5

monolingual and bilingual school-aged children. Typically developing children were chosen as a population of study since they constitute a first step in better understanding what is expected in school-aged children and may serve as foundation for future research studies of children with language or learning disorders. The project is set in five Canadian cities (Halifax, Moncton, Montréal, Ottawa, and Toronto) and data collection started with in-person procedures in 2016. In March 2020, an online data collection procedure was adapted. The present study describes the adapted procedures and compares the outcomes of online and in-person testing using discourse samples and standardized vocabulary testing for monolingual and bilingual children. The present study also examines whether the impact of modality differs across measures and what might account for these differences. Specifically, we asked:

1. Did discourse (conversation, expository, narration) or standardized vocabulary measures differ when testing was done in-person vs. online?
2. Did the impact of modality vary across productivity, proficiency, and syntactic complexity?
3. Did the impact of modality vary with age from 7 to 12 years of age?

Materials and methods

Participants

This study used a subset of data collected for a larger Canadian French/English Discourse study (FrEnDS-CAN) investigation of discourse development in school-age bilingual and monolingual children. For the present analyses, 127 (103 in-person, 24 online) English monolinguals and 78 (53 in-person, 25 online) simultaneous French-English bilinguals were included. The children were distributed across three age groups (7–8, 9–10, and 11–12 years). For the monolingual children, there were 57 female children (44 male and 26 who did not disclose) and the mean age was 8.57 (SD = 1.525). For the simultaneous bilingual children, there were 24 female children (24 male and 30 who did not disclose) and the mean age was 8.79 (SD = 1.598). For more information about how our participants were distributed across groups, please see Table 1 below. All children were typically developing, with no diagnosed or suspected language, hearing or learning difficulties (as established through parent report). They were recruited through their schools, through posters placed in public places and on social media.

The monolingual children were recruited through research teams in New Brunswick, Nova Scotia, Ontario and Quebec. They were exposed to only English in the home and attended English-language schools. Since Canada has two official languages, English-schooled children are required to take French as an academic subject (core French) in elementary school starting in grade 4 in both provinces. Thus, children were considered English monolinguals if they were exposed to French <10% of the time (e.g., only through these core classes at school), established through parent report.

The simultaneous bilingual children were recruited through research teams in Montréal (Quebec), Ottawa (Ontario), and Moncton (New Brunswick). They were exposed to both English and French from before the age of three and were able to complete testing in each language. In the home language questionnaire, there were 18 students who primarily spoke English at home, 19 who primarily spoke French at home, and 9 who reported an even split between the two languages. These children attended French-language schools and lived in communities where both French and English were spoken, so were likely to encounter both languages outside the home. In French-language schools, English classes are required starting in grades 1 or 2 in Quebec, grade 4 in Ontario, and grade 3 in New Brunswick. While these children were assessed in both languages, for the purposes of this study, we will only be discussing their English performance.

Procedures

Ethical approval was obtained through each participating university and school district. Consent forms were distributed

and collected through schools or *via* email. Parents confirmed their child's eligibility to participate by checking appropriate boxes on the consent form. Monolinguals were tested in a single session; bilinguals were tested in two sessions by different examiners: one in English, the other in French. Parallel tests and tasks were administered in French and English. The order of language testing was counterbalanced such that equal numbers of children within every age group were tested in French or English first. In each language, the test protocol began with a standardized vocabulary comprehension test. This was followed by conversational, narrative and expository language samples, in which the order of administration was also counterbalanced within age groups and within each language for bilinguals. The in-person, but not the online protocol, ended with the administration of a non-word repetition task. Since the non-word repetition task was not administered in both contexts, it is not discussed further. Any French language testing is also not included in this study and will not be discussed further. Examiners were graduate students, undergraduate students, or researchers. All student examiners were trained on test and task administration by the same Ph.D. student.

In-person testing occurred in a quiet area of the child's school or in a testing room in the research laboratories of participating universities. For in-person testing, sessions were recorded using a digital voice recorder which was placed next to the child. In March 2020, in-person testing was suspended in response to the COVID-19 pandemic. Consequently, materials were modified to accommodate online testing (discussed in the materials section). Zoom and Microsoft TEAMS platforms were used to test children online. The research team trained the testers in the use of the online platforms and the online administration of the full testing protocol was piloted on two children (9 and 10 years of age). These videotaped pilot sessions were then used as examples to train other testers. Children who were tested online used their home computer in a quiet area of their house. For online testing, sessions were recorded locally on the tester's computer using the platform of the parent's choice (either Zoom or Microsoft Teams). When technical issues arose which impeded the audio quality, testers would stop testing and troubleshoot the connection with the family until the audio quality was sufficient. A parent was asked to be available during testing in case the child experienced any technical or other difficulties. This usually meant they were in the room with the child but not sitting with them. Testers made notes of any significant behavioral or technical issues that arose during online testing using a common form.

Materials

Vocabulary comprehension

The *Peabody Picture Vocabulary Test-4* (*PPVT-4*; Dunn and Dunn, 2007) was used to test vocabulary comprehension in English. The *PPVT-4* has good reliability and validity and is commonly used for both educational and research purposes. In this vocabulary comprehension test, children point to one of four pictures in response to spoken word stimuli. Each test has a start point determined by age. Basals and ceilings were determined following manual instructions. In-person testing used the test booklet. Online testing followed the same procedures as in-person testing except that the stimuli were presented as images and the examiner's screen was shared with the child; each image presented a different item's picture stimuli.

Conversational samples

A conversational sample of at least 10 min was collected by the adult examiner following the *Systematic Analysis of Language Transcripts* (*SALT*; Miller and Chapman, 2012) interview protocol. This involved talking to the child about topics of interest to them. The child was initially asked what they would like to talk about and as the conversation progressed, additional topics from a common set (e.g., family, pets, school, hobbies) were introduced if needed. This flexible protocol was selected since it allowed children to choose a topic they were motivated to speak about, and therefore gave them an opportunity to demonstrate their oral language abilities. All examiners were instructed to listen as often as possible, and to only participate in the conversation when necessary to keep the child engaged. For example, examiners were instructed to ask open-ended questions such as "What were your favorite memories from your trip?" or "What do you like about art class?"

Narrative samples

Two narrative samples were collected in each language, the first using a story stem (setting information provided to generate a story) and the second using a single picture elicitation task. The story stem narratives are not analyzed here, so will not be discussed further. The single-picture elicitation was one of three narrative tasks included in the *Test of Narrative Language-2* (*TNL-2*; Gillam and Pearson, 2017; the revised test materials were shared with the research team prior to publication). It uses a 'give a story, get a story' format in which the examiner first tells a story about a complex picture (i.e., two children hiding behind a rock watching a treasure chest with either a dragon or pirates guarding it), asks 12 comprehension questions (6 literal, 6 inferential) about the treasure story, then produces a second picture (two children hiding as a family of aliens deboards from a spaceship or two children watching as an ogre has a Pegasus on a rope) and asks the child to produce "an even better" story using this new picture. Throughout this task, the child and the researcher were always able to see and reference the pictures (e.g., the pictures were either placed on a desk before the child or the researcher shared their screen with the pictures on it). The *TNL-2* is widely used and has high reliability and validity.

Expository samples

Expository samples were collected using an adaptation of "The Favorite Game or Sport Task" developed by Nippold et al. (2005) and modified by Heilmann and Malone (2014). In this protocol, children were asked to describe how to play a game or sport of their choice. Games without clear rules or an ending and videogames were excluded. Children first identified the game or sport they would describe. They were given a few minutes to plan what they wished to say. To encourage them to think broadly, the children were presented with eight areas they might discuss: What you try to do, Getting ready to play, Starting the game, How you play, Rules, Scoring, and Ending the game. During in-person testing, each of these components was printed on a two- by four-inch card and presented randomly in an array in front of the child with a brief verbal explanation of each (e.g., "you could talk about what you're trying to do in the game"). If asked, the examiner could re-read the cards; this occurred occasionally with younger children. When planning, the children could rearrange the cards as needed although only a minority of children did so. For online testing, circles with these same components printed in them were presented individually *via* screen sharing, until all were present on the screen where they remained for the duration of the task. One of ten pre-determined orders of presentation were used, selected randomly by the examiner. Once all the topic areas were introduced either on the computer screen or with cards, the child was given as much time as they wished to plan. When they said they were ready, they were asked to explain the game. When they indicated they were done, the examiner asked the child to explain any special strategies that could be used to win the game.

Analyses

Transcription

Each discourse sample was transcribed into Communication-units (C-units) by trained graduate students. A C-unit is defined as an independent clause and its modifiers (Hughes et al., 1997); it may be incomplete (e.g., a few words in response to a question). Transcriptions followed slightly modified *SALT* conventions. The modifications included: writing contractions as separate words rather than slashing them (e.g., don't = do not) and slashing the past participle –en (e.g., was give/en). Lexical verbs were identified using a [v] code next to the verb (e.g., was give/en[v]). Transcription of conversational samples began as soon as the child was interacting naturally and continued for 10 consecutive minutes. If there was not 10 min of conversation available, additional time was taken from conversations between child and examiner throughout the session to obtain the full 10 min. Narrative and expository transcriptions began after the instructions were completed and ended when the child indicated they were finished. Sample transcripts were saved in separate files and checked and corrected with reference to the session's audio-recording by a second, experienced transcriber.

Microstructure metrics

SALT software was used to generate microstructure metrics separately for each discourse file. For the purposes of this study, three microstructure metrics were computed. The first was a language proficiency metric, the mean length of utterance in morphemes (MLCUm). This is automatically calculated in *SALT* by averaging the number of morphemes (i.e., word roots and slashed morphemes) per C-unit. This study uses morphemes instead of words since that metric is more commonly utilized by clinicians and since this is not a cross-linguistics study. Secondly, we looked at a productivity metric, the number of total words produced by the child, referred to as NTW. Finally, we looked at syntactic complexity. In this case, we created a syntactic complexity score (SC) in SPSS by dividing the number of lexical verbs by the total number of C-units. These metrics were specifically chosen because they represent the three microstructure measures that might interest clinicians in online assessment: language proficiency, productivity, and syntactic complexity.

Design

To investigate the impact of modality on language assessment, we tested two groups of school-aged children in English: monolingual and simultaneous bilingual. Three age groups were included: 7–8-, 9–10-, and 11–12-year-olds. Three different language samples were collected from each participant: conversation, expository and narrative. Within each of these samples, we looked at three micro-structure measures (dependent variables): mean length of C-Unit in morphemes (MLCUm), number of total words (NTW) and syntactic complexity (SC). We also analyzed the raw scores of a standardized test of vocabulary (PPVT). The participants were tested either online or in-person.

Statistical analyses

All statistical analyses were run in SPSS version 28. Descriptive statistics (means, standard deviations) were generated on all English standardized test scores and discourse sample measures, separately for children tested in-person and online. Descriptive statistics were computed for the whole group of monolinguals and for the whole group of bilinguals as well as for each of the three age groups within those groups. We then completed two types of statistical analyses: ANOVAs for vocabulary analyses or MANOVAs for discourse analyses, and Bayesian *t*-tests.

A two-way modality (online vs. in-person) by age (7–8, 9–10, 11–12) between-subjects ANOVA tested mean differences

on raw PPVT scores, separately for monolingual and bilingual groups. Two-way modality (online vs. in-person) by age (7–8, 9–10, 11–12) between-subjects MANOVAs tested mean differences in the three discourse measures, separately for conversation, exposition, and narration for the two language groups (monolingual and bilingual). Preliminary analyses for the MANOVAs were completed. Significant main effects for age were examined with *post-hoc* comparisons with a Bonferroni correction for alpha level. Boxplots showed there were no outliers. The data were normally distributed, as assessed by Shapiro-Wilk's test of normality ($p > 0.05$) and there was homogeneity of variances ($p > 0.05$) and covariances ($p > 0.05$), as assessed by Levene's test of homogeneity of variances and Box's M test, respectively. Alpha was set at 0.05 a priori for all analyses.

Bayesian statistics were used to follow-up when non-significant modality main effects were obtained. The Bayes factor (BF_{01}) statistic was used. Bayes factors globally confirm that the absence of difference is not due to a lack of power but to the fact that the two modalities are equal (Brydges and Gaeta, 2019). A $BF_{01} > 1$ indicates evidence for the null hypothesis (H0). The further a value is from 1 (up to 100) the stronger the evidence is in favor of the null hypothesis (IBM, 2021; van Doorn et al., 2021).

Results

Monolinguals

Table 2 provides the means and standard deviations for PPVT scores and the conversation, expository, and narrative discourse measures. For a summary of the results, please see Table 1 in Appendix 1.

Vocabulary

The two-way ANOVA was conducted on receptive vocabulary raw scores. Only the main effect for age was significant, $F_{(2,105)} = 11.808$, $p < 0.001$, partial $\eta^2 = 0.191$. *Post-hoc* paired comparisons on the age effect found vocabulary raw scores differed significantly for all age groups and increased with increasing age (7–8: M = 141.57, SD = 16.70; 9–10: M = 157.92, SD = 18.31; 11–12, M = 176.26, SD = 16.74). The follow-up Bayesian analysis found a BF_{01} of 2.106, which provided only anecdotal evidence for the equivalence of a modality effect for raw receptive vocabulary scores in these monolingual children.

Conversational discourse

A two-way MANOVA analyzed MLCUm, NTW and SC in conversation. No significant main effects or interactions were obtained for MLCUm or SC. However, for NTW, the main effects of modality [$F_{(1,120)} = 5.579$, $p = 0.020$, partial $\eta^2 = 0.046$] and age [$F_{(2,120)} = 4.158$, $p = 0.018$, partial $\eta^2 = 0.067$] were significant, but not the interaction. In terms of modality, more words were produced in-person (M = 839.98; SD = 239.21) than on-line (M = 759.64; SD = 278.12) in conversations. *Post-hoc* paired comparisons showed the number of words produced in conversation increased significantly from 7 to 8 years (M = 721.24, SD = 228.98) to both 9–10 (M = 880.98, SD = 222.40) and 11–12 (M = 923.23, SD = 259.47) years of age. Bayesian statistics confirmed moderate evidence that the two modalities were equivalent for MLCUm ($BF_{01} = 5.14$) and SC ($BF_{01} = 5.51$).

Expository discourse

A two-way MANOVA analyzed MLCUm, NTW and SC in expository samples. For MLCUm, significant main effects were obtained for modality [$F_{(1,120)} = 5.406$, $p = 0.022$, partial $\eta^2 = 0.045$] and age [$F_{(2,120)} = 3.196$, $p = 0.045$, partial $\eta^2 = 0.053$]. Additionally, the main effect of age was significant for NTW [$F_{(2,120)} = 3.421$, $p = 0.036$, partial $\eta^2 = 0.056$]. No other main effects or interactions reached significance. The modality effect indicated greater MLCUm for in-person (M = 11.30, SD = 2.20) compared to on-line (M = 10.48, SD = 1.48) expository samples. *Post-hoc* paired comparisons of age groups indicated that the 11–12 group (M = 12.08, SD = 1.69) produced significantly longer C-units in expository samples than the 7–8 group (M = 10.49, SD = 2.33). Additionally, the 11–12 group produced significantly more words (M = 620.04, SD = 328.75) than the 7–8 age group (M = 378.02, SD = 247.58) in these samples.

Follow-up Bayesian analyses showed Bayesian factors of $BF_{01} = 3.59$ for NTW and $BF_{01} = 2.96$ for SC. These indicated the strength of evidence was moderate for NTW but anecdotal for SC that the two modalities were equivalent.

Narrative discourse

The narrative MANOVA revealed a main effect of age for SC [$F_{(2,120)} = 10.244$, $p < 0.001$, partial $\eta^2 = 0.153$]. No other main effects or interactions were obtained. *Post-hoc* paired comparisons of the age effect revealed that the youngest group (M =1.28, SD = 0.35) produced significantly fewer verbs per C-unit than either the middle (M = 1.57, SD = 0.37) or oldest (M = 1.63, SD = 0.32) age groups.

Bayesian follow-up analyses showed a $BF_{01} = 1.242$ for MLCUm, $BF_{01} = 3.337$ for NTW, and $BF_{01} = 2.375$ for SC. These scores provide moderate evidence of modality equivalence for NTW and anecdotal evidence for the other narrative discourse measures.

TABLE 2 Means and standard deviations of monolingual participants on all measures.

			7–8 years ($n = 53$)		9–10 years ($n = 48$)		11–12 years ($n = 26$)	
			In-person ($n = 48$)	Online ($n = 5$)	In-person ($n = 38$)	Online ($n = 10$)	In-person ($n = 17$)	Online ($n = 9$)
Receptive vocabulary (PPVT raw score)								
		Mean	140.8	149.0	157.09	160.9	177.9	171.7
		SD	17.03	12.27	18.43	18.64	16.03	19.41
Conversation								
	MLCUm	Mean	7.37	7.783	7.55	7.368	7.67	7.85
		SD	1.71	2.719	1.24	1.11	1.24	1.65
	NTW	Mean	729.4	629.5	909.8	762.7	980.8	814.4
		SD	233.4	307.3	206.5	258.5	224.8	298.2
	SC	Mean	0.905	1.00	0.924	0.815	0.926	0.956
		SD	0.231	0.357	0.177	0.138	0.184	0.282
Expository								
	MLCUm	Mean	10.50	10.37	11.68	9.89	12.56	11.18
		SD	2.41	1.64	1.72	1.53	1.76	1.17
	NTW	Mean	371.8	432.4	395.9	492.9	664.1	536.9
		SD	254.8	183.6	231.6	299.3	355.9	269.1
	SC	Mean	1.82	1.43	1.73	1.39	1.71	1.45
		SD	2.02	0.36	0.384	0.308	0.387	0.264
Narrative								
	MLCUm	Mean	8.71	9.1	9.46	10.33	9.94	10.68
		SD	2.37	1.34	2.77	1.72	1.30	2.91
	NTW	Mean	203.5	236.8	246.4	276.9	347.8	339.6
		SD	173.6	265.5	149.9	163.2	248	184
	SC	Mean	1.30	1.08	1.53	1.69	1.60	1.68
		SD	0.342	0.439	0.377	0.337	0.258	0.447

PPVT, Peabody Picture Vocabulary Test, 4th edition (Dunn and Dunn, 2007); MLCUm, mean length of communication unit in morphemes; NTW, number of total words; SC, syntax complexity; SD, standard deviation.

Simultaneous bilinguals

Only the English tasks were analyzed for the bilingual children to be congruent with the monolingual group. Table 3 shows the means and standard deviations for vocabulary (PPVT) scores and conversation, expository, and narrative discourse measures. For a summary of the results, please see Table 1 in Appendix 1.

Vocabulary

The two-way ANOVA revealed only a significant main effect of age for these bilingual children, $F_{(2,53)} = 18.52$, $p < 0.001$, partial $\eta^2 = 0.421$. *Post-hoc* paired comparisons showed that PPVT raw scores were significantly lower for the 7–8-year-olds (M = 124.8, SD = 24.95) than either the 9–10 (M = 157.70, SD = 26.85) or 11–12-year-olds (M = 176.07, SD = 14.82).

Bayesian follow-up analysis revealed a BF_{01} of 4.499. This provided moderate evidence for modality equivalence on receptive vocabulary.

Conversational discourse

The two-way MANOVA analyzing conversational discourse measures revealed significant main effects of age for MLCUm [$F_{(2,76)} = 6.156$, $p = 0.003$, partial $\eta^2 = 0.148$], NTW [$F_{(2,76)} = 7.156$, $p < 0.001$, partial $\eta^2 = 0.168$], and SC [$F_{(2,76)} = 3.861$, $p = 0.026$, partial $\eta^2 = 0.098$]. No other main effects or interactions were significant. All three measures increased with age. *Post-hoc* paired comparisons revealed significant improvement in MLCUm from 7 to 8 (M = 6.69, SD = 1.54) to 11–12 (M = 8.25, SD = 1.62). A similar pattern was found for the NTW (7–8: M = 635.07, SD = 257.94; 11–12: M = 907.43, SD = 256.23) and SC (7–8: M = 0.79, SD = 0.24; 11–12: M = 0.98, SD = 0.26) measures.

TABLE 3 Means and standard deviations of bilingual participants on all measures.

			7–8 years ($n = 29$)		9–10 years ($n = 28$)		11–12 years ($n = 21$)	
			In-person ($n = 19$)	Online ($n = 10$)	In-person ($n = 18$)	Online ($n = 10$)	In-person ($n = 16$)	Online ($n = 5$)
Receptive vocabulary (PPVT raw score)								
		Mean	125.5	123.4	152.6	165.6	174.2	183.0
		SD	27.26	21.9	26.8	26.45	13.33	21.17
Conversation								
	MLCUm	Mean	6.807	6.480	7.503	7.866	8.081	8.772
		SD	1.462	1.726	1.880	1.307	1.572	1.835
	NTW	Mean	672.8	563.2	702.6	840.1	904.1	918.2
		SD	273.3	221.1	240.2	152.4	247.6	313.4
	SC	Mean	0.807	0.751	0.831	0.875	0.976	1.001
		SD	0.255	0.212	0.243	0.114	0.239	0.337
Expository								
	MLCUm	Mean	9.774	9.671	11.16	10.21	11.87	11.28
		SD	2.273	1.398	4.428	1.648	2.106	3.626
	NTW	Mean	303.4	330.7	504.5	507.5	560.8	675.6
		SD	223.9	285.5	565.2	317.1	283.4	301.9
	SC	Mean	1.477	1.357	1.494	1.414	1.707	1.526
		SD	0.515	0.420	0.378	0.333	0.407	0.567
Narrative								
	MLCUm	Mean	8.213	8.708	9.562	10.46	9.744	10.75
		SD	1.450	1.756	2.360	1.958	1.380	2.142
	NTW	Mean	243.1	134.7	208.9	232.4	327.4	340.6
		SD	152.0	87.72	120.1	61.44	213.3	149.4
	SC	Mean	1.226	1.207	1.385	1.621	1.554	1.799
		SD	0.262	0.302	0.436	0.376	0.208	0.551

PPVT, Peabody Picture Vocabulary Test, 4th edition (Dunn and Dunn, 2007); MLCUm, mean length of communication unit in morphemes; NTW, number of total words; SC, syntax complexity; SD, standard deviation.

Bayesian results were: $BF_{01} = 5.340$ for MLCUm, $BF_{01} = 5.295$ for NTW, and $BF_{01} = 5.209$ for SC. Thus, moderate evidence supports the conclusion that the two modalities were equivalent for all of the conversational measures for the bilingual group.

Expository discourse

Two-way MANOVA results showed a main effect of age for MLCUm [$F_{(2,76)} = 3.513$, $p = 0.035$, partial $\eta^2 = 0.091$] and NTW [$F_{(2,76)} = 3.807$, $p = 0.027$, partial $\eta^2 = 0.098$] only. No other main effects or interactions were significant. *Post-hoc* paired comparisons showed lower MLCUm scores for 7–8 (M = 9.74, SD = 1.99) than 11–12 (M = 11.73, SD = 2.45) age groups as they were for NTW scores (7–8: M = 312.83, SD = 242.10; 11–12: M = 588.14, SD = 284.57).

Bayesian analyses resulted in a MLCUm BF_{01} of 4.43, a NTW BF_{01} of 5.12, and a SC BF_{01} of 2.42. Leading us to conclude there is moderate evidence (anecdotal for SC) for modality equivalence between groups on any expository measure.

Narrative discourse

Finally, a two-way MANOVA revealed main effects of age for all narrative measures: MLCUm [$F_{(2,76)} = 6.295$, $p = 0.003$, partial $\eta^2 = 0.152$], NTW [$F_{(2,76)} = 4.939$, $p = 0.010$, partial $\eta^2 = 0.124$], and SC [$F_{(2,76)} = 9.338$, $p < 0.001$, partial $\eta^2 = 0.211$]. No other main effects and interactions were significant. *Post-hoc* paired comparisons showed 7–8 MLCUm (M = 8.39, SD = 1.55) to be significantly lower than either 9–10 (M = 9.88, SD = 2.23) or 11–12 (M = 10.00, SD = 1.60) age groups. This was also true for SC (7–8: M = 1.22, SD = 0.27), 9–10: M = 1.47, SD = 0.42; 11–12: M = 1.62, SD = 0.33). In contrast, 7–8-year-olds (M = 204.39, SD = 140.96) produced fewer words (NTW) in narratives than 11–12-year-olds only (M = 330.70, SD = 195.55); and the Bayesian revealed anecdotal certainty that there were no significant effects of modality for

MLCUm (BF_{01} = 0.8720), and SC (BF_01 = 0.1126. However, for NTW, the BF_{01} was 3.577, indicating moderate evidence of modality equivalence.

Discussion

The current study examined the comparability of online and in-person assessment on conversational, expository, and narrative discourse across both monolingual and simultaneous bilingual speakers of English. Specifically, we looked at metrics of productivity, proficiency, and syntactic complexity across these three forms of discourse. We furthermore examined the effect of age against the previous two questions. Overall, our results indicated that most measures seem to be comparable across in-person and online assessment contexts. For the monolingual group, there were no differences due to modality on either vocabulary or narrative measures. However, there were two distinct differences due to modality for conversation. First, we saw an impact of modality on the productivity metric of the conversational measure in favor of the in-person group. Second, we saw an impact of modality on the proficiency and syntactic complexity measure of the expository measure in favor of the in-person group. Finally, while students improved with age, the effect of modality did not vary with the age of the participants.

For the simultaneous bilingual group, we saw no differences across vocabulary, conversation, expository or narrative. More specifically, we saw no differences in either productivity or syntactic complexity due to the assessment context. While we did see that students improved with age on these measures, age did not have an impact on the effect of assessment modality. It is interesting to note that, despite differing in exposure to and use of the target language compared to their monolingual counterparts, we did not see differences due to assessment modality for the simultaneous bilingual group. However, it should also be noted that the simultaneous bilingual group was smaller, which may have affected the power level of these analyses. The theoretical and clinical implications of these findings will be discussed below.

Across both language groups and all ages, we saw no differences on receptive vocabulary when comparing the in-person and online assessment groups. We speculate that this can be explained by several factors. First, and most importantly, the PPVT had already been adapted to online assessment and validated. It was already common practice to use an electronic version t of the PPVT instead of a paper copy, even in person, which was very simple to use in an online format. Furthermore, the PPVT can be more easily adapted for online assessment difficulties. For example, if children are shy or anxious about the session, they can either hold up the number of fingers to indicate which picture they choose, or they can type it in the chat. There is also less impact due to Wi-Fi or audio issues since both researchers and participants are only communicating single words (e.g., "Picking" from examiner and "two" from the child). We can therefore state with some degree of confidence that the results of receptive vocabulary tasks are comparable across in-person and online testing contexts.

For our conversational measure, we saw no differences across either language group on syntactic complexity. There were also no differences in modality due to the age group of the participant. However, in the monolingual group, we did see a significant difference between the online and in-person testing groups, in favor of the in-person group, on the number of total words (NTW), which is a measure of overall productivity. This may be attributed to several things. First, the context itself caused changes to the task that the researchers could not control. For example, the Wi-Fi and audio quality seemed to impact the conversational measure the most. Researchers often had to ask students to repeat themselves, which can cause students to withdraw. If students froze, they would lose their train of thought and have more difficulty getting their momentum back. It also seemed like students had more difficulty engaging with the task if they experienced interruptions. Furthermore, the conversational measure is the least structured of all the tasks. We suspect that the conversational measure shows differences between contexts on productivity because it is the only turn-taking task and is therefore more susceptible to factors such as lags in audio quality, lack of gestural language, less fluidity, and less non-verbal cues. For shy or anxious students, even in person, this can be a daunting task. Emerging research has shown that the current COVID-19 pandemic has exacerbated these challenges such as language anxiety (Imran et al., 2020; Lavigne-Cerván et al., 2021; Orgilés et al., 2021). In the clinical implications section, we will discuss how these challenges could be mitigated by future researchers.

For the expository task, conversely, we saw no differences across either language group on productivity. There were also no differences in modality due to the age group of the participant. However, in the monolingual group, there was a significant difference between the in-person and online testing groups, in favor of the in-person group, on the mean length of C-unit in morphemes (MLCUm), which is a measure of syntactic complexity and language proficiency. This was an interesting finding since the expository task is highly structured, and the most likely to produce complex language. However, we speculate that this difference may be due to the adaptation of the task to the online format. To the best of our knowledge, the "Favorite game or sport task" has not been administered online in a research study previously. While some aspects of this task were easy to adapt (e.g., asking the questions about their favorite sport), others were more complex. In the in-person version of this task, students are provided with optional "prompt" cards in a randomized order, which they can refer to, use as a physical manipulative, or ignore entirely. In the online version, we created several randomized versions of these cards which would appear on the screen in front of the child. However, their

positionality on the screen seemed to give the cards more weight, and almost all of the children used them throughout this task. This could, in turn, make the task more formulaic since the students were simply responding to each prompt individually. In person, children were more expansive and creative with their descriptions. Another possibility is the impact of cognitive load. Expository discourse is already a complex task that requires organization, explicit instructions, and complex language. It is possible that children struggle to complete more complex tasks online since some of their cognitive load is focused on the online testing format in addition to the assessments themselves. In summary, we suspect that the syntactic complexity of the expository measure was lower online since it is the most challenging of the tasks. The high cognitive load, combined with the challenging online interaction, may have posed particular challenges. In the clinical implications section, we will discuss how this challenge could also be mitigated by future researchers.

Across both language groups and all ages, we saw no differences on the productive narrative task when comparing the in-person and online assessment groups. We speculate that this can be explained by several factors. While this task has not been previously adapted in a research study, it was simpler to move to an online format. Children would simply see the picture on their screens instead of on the desk in front of them, and the prompts were otherwise identical. This task also usually feels less like a "test" to students, and it seems easier for them to engage since it is highly imaginative. While there were also audio and Wi-Fi issues during this task, the participants seemed to lose their train of thought less since they were continuing a narrative. We can therefore state with some degree of confidence that the results of productive narrative tasks are comparable across in-person and online testing contexts.

Clinical implications

The current COVID-19 pandemic has influenced the practices of researchers and clinicians within the healthcare field, including speech-language pathologists. Historically, speech-language pathology has depended largely on in-person interactions to assess children's language abilities. However, since the beginning of the COVID-19 pandemic, the use of online assessment and treatment models have become widely implemented, often being offered as the primary method of service. For researchers and clinicians, it is important to be aware of and account for any differences that may result from assessment modality. Particularly for clinicians, the transformations of these in-person interactions to an online medium must consider which important insights can be captured during language sample analysis. Similar to previously discussed research (Taylor et al., 2014; Manning et al., 2020), the aforementioned results further support using language sample analyses gathered online for various discourse types (i.e., conversation, narrative, and expository language) for both monolingual and bilingual speakers aged 7–12 years. Both modalities can provide researchers and clinicians with accurate and reliable information about the child and their language abilities.

However, it is also important to discuss the modifications that may need to be made as indicated by the results of this study. As previously stated, it appears that receptive vocabulary and narrative measures are more easily adaptable to the online assessment context. To successfully use a conversational assessment, we would make two recommendations. First, it is crucial that clinicians and researchers thoroughly test any Wi-Fi and audio issues, and only proceed with the assessment if a minimum threshold is met. Furthermore, it would be advantageous to ensure the child is calm, comfortable, and engaged prior to beginning the conversational task. This could mean having a relative sit with them, informally chatting before starting the task, asking guardians for topics ahead of time, etc. Otherwise, the conversational task may not be as reflective of the students' language abilities as it would be in person. For the expository task, we speculate that the "Favorite game or sport" task may be more difficult to administer online. An expository task that does not rely on prompt cards might be more suitable for online assessment (e.g., explaining how to make a peanut butter and jelly sandwich). Alternatively, prompt cards could be eliminated for all modalities. Finally, clinicians may notice qualitative differences in online testing, including interruptions due to technical errors, lack of tactile information, or interjections from family members.

Limitations and future directions

Research comparing in-person and online assessment is relatively new in the field of speech-language pathology. The current study found that monolingual children differed on conversational productivity and expository syntactic complexity. Future studies may want to investigate why MLCUm was more sensitive to differences than other syntactic complexity measures (the number of lexical verbs by the number of C-units). Our research focused on three discourse measures for typically developing children aged 7 to 12 years of age. However, the distribution across conditions was uneven, with a higher number of students in person than those who were tested online. Future studies may want to expand on our results and focus on other populations or discourse measures to further confirm if online and in-person assessment can be used interchangeably. For example, future studies may want to include a sample of younger children, or children with language or learning difficulties, such as a Developmental Language Disorder. Our results comparing in-person and online testing may vary for other populations of children. This is especially true for the expository and conversational measures, where any

differences in modality may be exaggerated in non-typically developing populations. Furthermore, the current findings should be validated and supplemented by studies using a within-subject design. Additionally, future studies could include independent measures of the children's language and cognitive abilities. In the same vein, future studies may want to look at other microstructure or macrostructure measures. Finally, more research is needed on the impact of bilingualism on assessment modality. While no differences were found in this paper, it would be beneficial to replicate these results with sequential bilingual children and bilingual students from other language backgrounds.

Conclusion

In conclusion, it is possible to conduct measures of discourse online with similar results to that of in-person language sampling data for monolingual and bilingual children aged 7–12 years. The evidence of this study suggests that receptive vocabulary and narrative measures are reliable assessments to be used in an online context. Conversational measures may be comparable with the aforementioned safeguards in place (e.g., audio, Wi-Fi, situating the child). Expository measures should be used with caution until further research has explored the differences between modalities. While pivoting to online services can be difficult for researchers and clinicians, language sampling is a valuable resource and requires little materials for both monolingual and bilingual children. Based on the results of the current study, researchers and clinicians can feel confident in continuing to use language sampling as an informative assessment tool in the provision of online services.

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 Dalhousie University Research Ethics Board: Health Sciences. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

DB MB, and EK-RB were responsible for the introduction. DB and EK-RB were responsible for materials and methods. DB, SR, XC, BS, and EK-RB were responsible for the results. DB, VB, KB, and JM were responsible for the discussion. All authors were involved in the revision and DB was responsible for all final edits. All authors were involved in the design of this study, in addition to the acquisition, analysis, and interpretation of data for the work. Furthermore, all authors were involved in drafting the work and revising it critically for important intellectual content. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by an Insight Grant (#435-2016-1026) provided by the Social Sciences and Humanities Research Council (SSHRC) of Canada.

Acknowledgments

We would like to acknowledge the contributions of all members of the FrEnDs-CAN group for their unwavering support. We would also like to thank the school boards, principals, teachers, students, and families who made this study possible. We would like to thank all of the supporting institutions as well as SSHRC for their support of this study.

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

References

Berman, R., and Slobin, D. (1994). *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Berman, R. A., and Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: a developmental paradox. *Discourse Process* 43, 79–120. doi: 10.1080/01638530709336894

Bliss, L. S. (2002). *Discourse Impairments: Assessment and Intervention Applications*. Boston: Allyn and Bacon.

Brydges, C. R., and Gaeta, L. (2019). An introduction to calculating Bayes factors in JASP for speech, language, and hearing research. *J. Speech Lang. Hear. Res.* 62, 4523–4533. doi: 10.1044/2019_JSLHR-H-19-0183

Cabell, S. Q., Justice, L. M., Piasta, S. B., Curenton, S. M., Wiggins, A., Turnbull, K. P., et al. (2011). The impact of teacher responsivity education on preschoolers' language and literacy skills. *Am. J. Speech-Lang. Pathol.* 20, 315–330. doi: 10.1037/e584752012-090

Chenneville, T., and Schwartz-Mette, R. (2020). Ethical considerations for psychologists in the time of COVID-19. *Am. Psychol.* 75, 644. doi: 10.1037/amp0000661

Craig, H. K., Washington, J. A., and Thompson-Porter, C. (1998). Average C-unit lengths in the discourse of African American children from low-income, urban homes. *J. Speech Lang. Hear. Res.* 41, 433–444. doi: 10.1044/jslhr.4102.433

Cummins, J. (2000). *BICS and CALP. Encyclopedia of Language Teaching and Learning*, 76–79. Clevedon: Multilingual Matters.

Dewaele, J. M., and Pavlenko, A. (2003). "Productivity and lexical diversity in native and non-native speech: A study of cross-cultural effects", in *The Effects of the Second Language on the First*, eds V. Cook (Clevedon: Multilingual Matters). 120–141. doi: 10.21832/9781853596346-009

d'Orville, H. (2020). COVID-19 causes unprecedented educational disruption: is there a road towards a new normal?. *Prospects* 49, 11–15. doi: 10.1007/s11125-020-09475-0

Dunn, L., and Dunn, D. (1997). *Peabody Picture Vocabulary Test-3rd Edition*. Circle Pines, MN: American Guidance Systems. doi: 10.1037/t15145-000

Dunn, L., and Dunn, L. (1981). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., and Dunn, D. M. (2007). *PPVT-4: Peabody Picture Vocabulary Test*. Bloomington, MN: Pearson Assessments. doi: 10.1037/t15144-000

Eisenberg, A. R. (1985). Learning to describe past experiences in conversation. *Discourse Process* 8, 177–204. doi: 10.1080/01638538509544613

Eisenberg, S. L., Fersko, T. M., and Lundgren, C. (2001). The use of MLU for identifying language impairment in preschool children. *Am. J. Speech Lang. Pathol.* 10, 323–342. doi: 10.1044/1058-0360(2001/028)

Eriks-Brophy, A., Quittenbaum, J., Anderson, D., and Nelson, T. (2008). Part of the problem or part of the solution? Communication assessments of Aboriginal children residing in remote communities using videoconferencing. *Clin. Ling. Phonetics* 22, 589–609. doi: 10.1080/02699200802221737

Fong, R., Tsai, C. F., and Yiu, O. Y. (2021). The implementation of telepractice in speech language pathology in Hong Kong during the COVID-19 pandemic. *Telemed. E Health* 27, 30–38. doi: 10.1089/tmj.2020.0223

Gillam, R. B., and Pearson, N. A. (2017). *Test of Narrative Language–Second Edition (TNL-2)*. Austin, TX: Pro-Ed.

Guiberson, M., Rodríguez, B. L., and Zajacova, A. (2015). Accuracy of telehealth-administered measures to screen language in Spanish-speaking preschoolers. *Telemed. e-Health.* 21, 714–720.

Haaf, R., Duncan, B., Skarakis-Doyle, E., Carew, M., and Kapitan, P. (1999). Computer-based language assessment software: the effects of presentation and response format. *Lang. Speech Hear. Serv. Sch.* 30, 68–74. doi: 10.1044/0161-1461.3001.68

Heilmann, J., and Malone, T. O. (2014). The rules of the game: Properties of a database of expository language samples. *Lang. Speech Hear. Serv. Sch.* 45, 277–290. doi: 10.1044/2014_LSHSS-13-0050

Heilmann, J. J., Miller, J. F., and Nockerts, A. (2010). Using language sample databases. *Lang. Speech Hear. Serv. Schools* 41, 84–95. doi: 10.1044/0161-1461(2009/08-0075)

Hoff, E. (2009). "Language development at an early age: Learning mechanisms and outcomes from birth to five years," in *Encyclopedia on early childhood development*, 1–5.

Hughes, D., McGillivray, L. R., and Schmidek, M. (1997). *Guide to Narrative Language: Procedures for Assessment*. Eau Claire, WI: Thinking Publications.

IBM (2021). *Bayesian independent sample inference*. Available online at: https://www.ibm.com/docs/ro/spss-statistics/25.0.0?topic=statistics-bayesian-independent-sample-inference#fnsrc_1 (accessed March 06, 2022).

Imran, N., Zeshan, M., and Pervaiz, Z. (2020). Mental health considerations for children and adolescents in COVID-19 Pandemic. *Pak. J. Med. Sci.* 36, S67. doi: 10.12669/pjms.36.COVID19-S4.2759

Jia, G., Chen, J., Kim, H., Chan, P. S., and Jeung, C. (2014). Bilingual lexical skills of school-age children with Chinese and Korean heritage languages in the United States. *Int. J. Behav. Dev.* 38, 350–358. doi: 10.1177/0165025414533224

Kay-Raining Bird, E., Joshi, N., and Cleave, P. L. (2016). Assessing the reliability and use of the expository scoring scheme as a measure of developmental change in monolingual English and bilingual French/English children. *Lang. Speech Hear. Serv. Sch.* 47, 297–312. doi: 10.1044/2016_LSHSS-15-0029

Lavigne-Cerván, R., Costa-López, B., Juárez-Ruiz de Mier, R., Real-Fernández, M., Sánchez-Muñoz de León, M., and Navarro-Soria, I. (2021). Consequences of COVID-19 confinement on anxiety, sleep and executive functions of children and adolescents in Spain. *Front. Psychol.* 12, 334. doi: 10.3389/fpsyg.2021.565516

Le Normand, M. T., Parisse, C., and Cohen, H. (2008). Lexical diversity and productivity in French preschoolers: developmental, gender and sociocultural factors. *Clin. Ling. Phonet.* 22, 47–58. doi: 10.1080/02699200701669945

Leadholm, B., and Miller, J. F. (1992). *Language Sample Analysis: The Wisconsin Guide*. Madison, WI: Department of Public Instruction.

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.

Mansuri, B., Tohidast, S. A., Mokhlesin, M., Choubineh, M., Zarei, M., Bagheri, R., et al. (2021). Telepractice among speech and language pathologists: a KAP study during COVID-19 pandemic. *Speech Lang. Hear.* 1–8. doi: 10.1080/2050571X.2021.1976550. [Epub ahead of print].

Manzanares, B., and Kan, P. F. (2014). assessing children's language skills at a distance: does it work? *Perspect. Augment. Altern. Commun.* 23, 34–41. doi: 10.1044/aac23.1.34

Merriam-Webster. (n.d.). *Discourse. In Merriam-Webster.com dictionary*. Available online at: https://www.merriam-webster.com/dictionary/discourse (accessed October 19, 2021).

Miller, J., and Chapman, R. (2012). *Systematic Analysis of Language Transcripts (SALT)[Computer software]*. Middleton, WI: SALT Software LLC.

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: 10.1111/j.1540-5826.2006.00205.x

Nippold, M. A. (2009). School-age children talk about chess: does knowledge drive syntactic complexity? *J. Speech Lang. Hear. Res.* 52, 856–871. doi: 10.1044/1092-4388(2009/08-0094)

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., and MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: examining the use of complex syntax. *J. Speech Lang. Hear. Res.* 57, 876–886. doi: 10.1044/1092-4388(2013/13-0097)

Nippold, M. A., Hesketh, L. J., Duthie, J. K., and Mansfield, T. C. (2005). Conversational versus expository discourse: a study of syntactic development in children, adolescents, and adults. *J. Speech Lang. Hear. Res.* 48, 1048–1064. doi: 10.1044/1092-4388(2005/073)

Nippold, M. A., Mansfield, T. C., and Billow, J. L. (2007). Peer conflict explanations in children, adolescents, and adults: examining the development of complex syntax. *Am. J. Speech Lang. Pathol.* 16, 179–188. doi: 10.1044/1058-0360(2007/022)

Nippold, M. A., and Sun, L. (2010). Expository writing in children and adolescents: a classroom assessment tool. *Perspect. Lang. Learn. Educ.* 17, 100–107. doi: 10.1044/lle17.3.100

Orgilés, M., Espada, J. P., Delvecchio, E., Francisco, R., Mazzeschi, C., Pedro, M., et al. (2021). Anxiety and depressive symptoms in children and adolescents during covid-19 pandemic: a transcultural approach. *Psicothema* 33, 125–130. doi: 10.7334/psicothema2020.287

Owens, R. (2012). *Language Development: An Introduction, 8th Edn*. Boston, MA: AUyn and Bacon.

Paul, R., and Norbury, C. F. (2012). Language disorders from infancy through adolescence. St Louis, MI: Elsevier Health Sciences.

Putri, R. S., Purwanto, A., Pramono, R., Asbari, M., Wijayanti, L. M., and Hyun, C. C. (2020). Impact of the COVID-19 pandemic on online home learning: an explorative study of primary schools in Indonesia. *Int. J. Adv. Sci. Technol.* 29, 4809–4818. Available online at: http://sersc.org/journals/index.php/IJAST/article/view/13867

Reimers, F., Schleicher, A., Saavedra, J., and Tuominen, S. (2020). Supporting the continuation of teaching and learning during the COVID-19 pandemic. *OECD* 1, 1–38. Available online at: https://globaled.gse.harvard.edu/files/geii/files/supporting_the_continuation_of_teaching.pdf

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., and Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *J. Speech Lang. Hear. Res.* 52, 333–349. doi: 10.1044/1092-4388(2009/08-0183)

Schneider, P., Dubé, R. V., and Hayward, D. (2004). *The Edmonton Narrative Norms Instrument*. Available online at: http://www.rehabmed.ualberta.ca/spa/enni/ (accessed October 19, 2021).

Stadler, M. A., and Ward, G. C. (2005). Supporting the narrative development of young children. *Early Childh. Educ. J.* 33, 73–80. doi: 10.1007/s10643-005-0024-4

Taylor, O. D., Armfield, N. R., Dodrill, P., and Smith, A. C. (2014). A review of the efficacy and effectiveness of using telehealth for paediatric speech and language assessment. *J. Telemed. Telecare* 20, 405–412. doi: 10.1177/1357633X14552388

Turkstra, L. S., Quinn-Padron, M., Johnson, J. E., Workinger, M. S., and Antoniotti, N. (2012). In-person versus telehealth assessment of discourse ability in adults with traumatic brain injury. *J. Head Trauma Rehabil.* 27, 424.

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., et al. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychon. Bull. Rev.* 28, 813–826. doi: 10.3758/s13423-020-01798-5

Verhoeven, L., and Strömqvist, S. (Eds.). (2001). *Narrative Development in a Multilingual Context*, Vol. 23. Amsterdam: John Benjamins Publishing. doi: 10.1075/sibil.23

Westerveld, M. F., and Moran, C. A. (2013). Spoken expository discourse of children and adolescents: retelling versus generation. *Clin. Ling. Phonet.* 27, 720–734. doi: 10.3109/02699206.2013.802016

Zebib, R., Tuller, L., Hamann, C., Abed Ibrahim, L., and Prévost, P. (2020). Syntactic complexity and verbal working memory in bilingual children with and without developmental language disorder. *First Lang.* 40, 461–484. doi: 10.1177/0142723719888372

Appendix

TABLE A1 Summary of all results by population, measure and analysis.

Population, measure, and analysis	Dependent variables	Results
Monolinguals		
Vocabulary, ANOVA	PPVT Raw score	Modality, NS
		Age*
		Modality × Age, NS
Conversation, MANOVA	MLCUm	Modality, NS
		Age, NS
		Modality × Age, NS
	NTW	Modality*
		Age*
		Modality × Age, NS
	SC	Modality, NS
		Age, NS
		Modality × Age, NS
Expository, MANOVA	MLCUm	Modality*
		Age*
		Modality × Age, NS
	NTW	Modality, NS
		Age*
		Modality × Age, NS
	SC	Modality, NS
		Age, NS
		Modality × Age, NS
Narration, MANOVA	MLCUm	Modality, NS
		Age, NS
		Modality × Age, NS
	NTW	Modality, NS
		Age, NS
		Modality × Age, NS
	SC	Modality, NS
		Age*
		Modality × Age, NS
Simultaneous bilinguals		
Vocabulary, ANOVA	PPVT Raw score	Modality, NS
		Age*
		Modality × Age, NS
Conversation, MANOVA	MLCUm	Modality, NS
		Age*
		Modality × Age, NS
	NTW	Modality, NS
		Age*
		Modality × Age, NS
	SC	Modality, NS
		Age*
		Modality × Age, NS
Expository, MANOVA	MLCUm	Modality, NS
		Age*
		Modality × Age, NS
	NTW	Modality, NS
		Age*
		Modality × Age, NS
	SC	Modality, NS
		Age, NS
		Modality × Age, NS
Narrative, MANOVA	MLCUm	Modality, NS
		Age*
		Modality × Age, NS
	NTW	Modality, NS
		Age*
		Modality × Age, NS
	SC	Modality, NS
		Age*
		Modality × Age, NS

PPVT, Peabody Picture Vocabulary Test, 4th edition (Dunn and Dunn, 2007); MLCUm, mean length of communication unit in morphemes; NTW, number of total words; SC, syntax complexity; NS, not significant.
*Denotes statistical significance.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 26 July 2022
DOI 10.3389/fpsyg.2022.916214

OPEN ACCESS

EDITED BY
Wenchun Yang,
Leibniz Center for General Linguistics (ZAS), Germany

REVIEWED BY
Maria Garraffa,
University of East Anglia, United Kingdom
Ciara O'Toole,
University College Cork, Ireland

*CORRESPONDENCE
Stanislava Antonijevic
stanislava.antonijevic@nuigalway.ie

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

RECEIVED 08 April 2022
ACCEPTED 04 July 2022
PUBLISHED 26 July 2022

CITATION
Antonijevic S, Colleran S, Kerr C and Ní Mhíocháin T (2022) Online assessment of narrative macrostructure in adult Irish-English multilinguals. *Front. Psychol.* 13:916214. doi: 10.3389/fpsyg.2022.916214

Online assessment of narrative macrostructure in adult Irish-English multilinguals

Stanislava Antonijevic*, Sarah Colleran†, Clodagh Kerr† and Treasa Ní Mhíocháin†

Discipline of Speech and Language Therapy, School of Health Sciences, National University of Ireland Galway, Galway, Ireland

Background: Online assessment of narrative production and comprehension became an important component of language assessment during the COVID-19 pandemic. This study aimed to establish quantitative measures of narrative macrostructure in the production and comprehension of adult Irish-English bilinguals in an online assessment.

Methods: A total of 30 Irish-English bilingual adults participated in an online assessment of oral narrative production and comprehension. Narratives were elicited using LITMUS-MAIN for Irish and English. Story-tell elicitation method was used for all stories. Twenty participants produced Baby Birds and Baby Goats story pairs while 10 participants produced Cat and Dog story pairs. Quantitative measures of story structure, comprehension score, and the overall number of Internal State Terms (ISTs) in production and comprehension were compared across the story pairs, languages, and the output type (production vs. comprehension).

Results: A general linear model indicated no differences in either story structure or story comprehension scores across languages for both sets of stories. Combined analysis for all participants and stories indicated no difference in the story structure scores or comprehension scores across the languages or the story pairs. While the overall number of ISTs was the same across languages, a higher number of ISTs was observed in comprehension relative to production in both languages for Cat and Dog story pair only, but not for Baby birds and Baby goats' stories. The major benefit of using online assessment was the accessibility of participants. The major drawback was the inability to control the environment and the quality of the internet connection.

Conclusion and implications: While online assessment increased the availability of participants, which is a significant factor in rural Ireland characterized by low population density and the high percentage of Irish speakers, the availability of stable internet connection limited the applicability of online assessment. Measures of narrative macrostructure were stable across the languages and the story pairs. This is important because of high variability in exposure to Irish, frequent code-switching, and a high number of morphosyntactic errors due to rapid language change that characterizes

Irish-English bilinguals. Identifying reliable measures of language performance for Irish-English adult speakers is an important step toward establishing developmental norms for Irish-English bilinguals.

KEYWORDS

macrostructure, narrative, story grammar, Irish, Multilingual Assessment Instrument for Narratives MAIN, narrative production, narrative comprehension, internal state terms

Introduction

While online language assessment has been present in speech and language therapy/pathology for a while, telepractice came into focus recently due to the interruption of in-person assessments at the start of the COVID-19 pandemic. The restrictions in conducting in-person speech and language assessments created an urgent need to validate online assessments and assessment protocols. Based on several studies that compared speech and language assessments online and in-person, Peña and Sutherland (2022) concluded that it is possible to reliably assess children using online procedures. Especially relevant here is the study by Pratt et al. (2022) who compared online and in-person narrative comprehension using the Multilingual Assessment Instrument for Narratives (MAIN; Gagraina et al., 2019) in English and Spanish. The study indicated a high correlation between the scores reflecting comprehension of story macrostructure for online and in-person assessments. This is an encouraging finding and calls for further research into using the MAIN (Gagarina et al., 2019) as an online assessment of narrative production and comprehension.

The current study used the MAIN in Irish (O Malley, 2019) and English (Gagarina et al., 2012, 2019) to assess narrative production and comprehension in Irish-English multilingual adults. Irish (*Gaeilge*) is one of the three official languages but at the same time a minority language of the Republic of Ireland. During the last 30 years, a significant decline has been noted in the use of Irish in the homes and wider communities resulting in almost universal bilingualism with English. Rapid language change is a consequence of the close contact with English and includes changes in the use of phonology, morphology, and syntax; increased use of direct translations from English; and frequent code-switching with English (Ó Catháin, 2016; Nic Fhlannchadha and Hickey, 2019, 2021). Due to the rapid language change, it can be challenging to judge grammatical accuracy and decide what are acceptable morphosyntactic forms in the current use of Irish. This further constraints the use of language assessments focusing on morphosyntax such as sentence repetition tasks or assessment of narrative microstructure (Antonijevic et al., 2017, 2020). Instead, a narrative assessment focusing on macrostructure can be a more ecologically valid and more reliable language assessment for Irish. To support language assessment through Irish in speech and language therapy/pathology, the MAIN (Gagarina et al., 2019) was adapted to Irish (O Malley, 2019).

Narratives offer a less biased method of assessing language in bi/multilinguals than norm-referenced, standardized assessments because their structural aspects are shared across languages (Paradis et al., 2010; Peña et al., 2014; Boerma et al., 2016). In addition, narratives include the interpretation of knowledge beyond the specifics of a particular language (Gagarina et al., 2012). Narratives can be analyzed at the levels of microstructure and macrostructure (Gagarina et al., 2015). The microstructure is specific to individual languages as it refers to the lexical and grammatical elements used to form coherent narratives (Boerma et al., 2016; Bohnacker, 2016). Macrostructure refers to the global organization of the story that is fairly similar across languages (Gagarina et al., 2019). The current study focuses on the MAIN (Gagarina et al., 2012, 2015, 2019), a narrative assessment that was specifically developed for multilingual children from diverse linguistic and cultural backgrounds as one of the assessments in the LITMUS battery created within the Cost Action IS0804 "Language Impairment in Multilingual Society: Linguistic Patterns and the Road to Assessment." The original hypothesis of the group was that the story grammar knowledge as reflected in the narrative macrostructure would be invariant across languages of multilinguals (Gagarina et al., 2015). The MAIN includes four parallel stories comparable in the storyline, characters, and the number and structure of the episodes. The macrostructure of each story includes three full episodes depicted across six pictures. Episodes contain three core components: Goal (G), the objective of the main character; Attempt (A), their action aimed at achieving the goal; and Outcome (O), the result of the action. Episodes are framed with two Internal State Terms (ISTs). The initiating IST refers to the main character's emotional or cognitive state that initiates the setting of the goal and the attempt to achieve the goal. Closing IST is a reaction to the outcome of the action aimed at achieving the goal. Therefore, the structure of a full episode can be represented

as $IST_{(initiating)}$-GAO-$IST_{(reaction)}$. The important difference between these different structural components is that characters' actions and the outcomes of those actions are explicitly depicted while their goals and internal states need to be inferred by the narrator from the elicitation pictures (Gagarina et al., 2012, 2015, 2019). The same structure in all four stories enables comparison of the assessment scores across languages and different elicitation modes (tell, retell and tell after a model story). It also allows for pre- and post-assessment without the risk of training effects (Pesco and Kay-Raining Bird, 2016), and ensures that any differences in language performance are not caused by variations in task difficulty (Kapalková et al., 2016).

Comprehension of narrative macrostructure is assessed through ten questions referring to the Goal of the main character, their Attempt to achieve the goal, the Outcome, and the two ISTs of each episode. Questions related to ISTs provide information on the child's metalinguistic and metacognitive knowledge (Armon-Lotem et al., 2015), their comprehension of the plot, and their ability to interpret and explain the perspectives and intentions of the protagonists (Curenton and Justice, 2004; Nippold et al., 2005; Heilmann et al., 2010).

Several studies compared macro and microstructure in the narrative productions of multilingual children. For example, Hipfner-Boucher et al. (2015) examined macrostructure and microstructure in the narrative retelling of 4–6 years old typically developing English Language Learners (ELL) with different home languages and compared those to monolingual English peers in Canada. All the ELL children had the same average exposure to English in their educational settings but had either a high or low exposure to English at home. With respect to microstructure, the low English-at-home group had significantly lower scores for sentence length, vocabulary, and grammaticality than the monolinguals and the high English-at-home group. However, both groups of ELL children produced story grammar of similar complexity to their monolingual peers. The differences in exposure to the dominant language at home influenced microstructure but not macrostructure in ELL early school-age children (Hipfner-Boucher et al., 2015). Narrative microstructure and macrostructure were also compared across languages in typically developing simultaneous Norwegian-Russian bilingual children (age 4–5 years) by Rodina (2017). Using the MAIN narrative assessment in tell mode, the study indicated that macrostructure was comparable across the two languages in both production and comprehension while microstructure was sensitive to language exposure. For the dominant language Norwegian, when compared to the monolingual peers, narratives of the bilingual children did not differ in either microstructure or macrostructure. However, for the minority language Russian when compared to the monolingual peers, narratives of the bilinguals differed in all microstructure measures while there was no difference in the macrostructure measure of story complexity. Similar results were obtained for balanced Polish-English bilinguals (age 5–7 years) living in the UK and attending education in English. Using the MAIN narrative assessment in both tell and retell mode Otwinowska et al. (2020) found that children's performance was comparable across the languages on all macrostructure measures while differences between productions in Polish and English were observed in basic lexical and syntactic measures which refer to microstructure.

Further comparison of microstructure and macrostructure in the narratives of bilingual children indicated that macrostructure might better discriminate between typically developing (TD) children and children with developmental language disorder (DLD). Narrative macrostructure was compared across TD children and children with language impairment (LI) in Dutch monolinguals and bilinguals (aged 5–6 years) by Boerma et al. (2016). Using the MAIN narrative assessment in the "tell after a different model story" mode of elicitation, the study indicated that macrostructure measures did not differ across monolingual and bilingual TD groups while, at the same time, macrostructure scores reliably differentiated between TD and LI groups in both monolinguals and bilinguals. Given that the current study is the first step toward the final aim of using MAIN Gaeilge (Irish) as a clinical language assessment tool, these findings point toward the advantage of macrostructure scores for this purpose.

The studies comparing microstructure and macrostructure for different language pairs indicated similar macrostructure across the languages in multilingual children. This is in line with the previous findings and theoretical assumptions suggesting that narrative macrostructure relies on children's cognitive development including general information processing skills such as working memory, attention, and executive function related skills of organization and planning (e.g., Berman and Slobin, 1994; Friend and Bates, 2014). It is further proposed that children may transfer domain general conceptual base across the languages resulting in equivalent narrative macrostructure for all their languages (Cummins, 1979; MacWhinney, 2005). Comparable macrostructure across languages of bilingual children makes narrative assessment a potentially useful tool for language assessment of Irish speakers. Because of the near-universal bilingualism and the variability in language exposure to Irish and English, a measure that is potentially equivalent across the two languages and at the same time can differentiate between TD children and children with DLD would be an ideal assessment tool for the population of Irish-English bilinguals.

Most studies comparing macrostructure across the languages of multilingual speakers focused on children. The study by Gagarina et al. (2019) compared narrative macrostructure production in German, Russian, and Swedish in monolingual adults. The aim of the study was to provide benchmark data from monolingual adults for story structure (sum of the core story elements G, A, O, and ISTs produced in the narrative) and story complexity (combinations of the core elements G, A, and O within each episode) and to compare

those across the three languages. The MAIN Baby Goats and Baby Birds stories were used to elicit narratives employing the tell mode of elicitation. The story structure scores were similar across the languages, indicating that the elicitation pictures are cross-linguistically and cross-culturally robust. Adults did not show the ceiling effect and achieved relatively low scores for story structure with an average of 11–12 points out of a maximum of 17. When comparing the story structure scores for each story, Baby Goats' scores were slightly higher than Baby Birds' scores. This finding is consistent with the findings of Lindgren (2019) where higher scores for story comprehension were also found in Baby Goats' story. Narrative comprehension was not reported in this study. The findings provide important information about adults' production of narrative macrostructure and benchmark data for the MAIN story structure and story complexity in monolingual German, Russian, and Swedish speakers.

The current study contributes to the existing research by reporting data on narrative macrostructure production and comprehension by Irish-English bilingual adults and establishing a baseline for macrostructure measures in this population. Given the near universal bilingualism with English, it would be impossible to benchmark the narrative assessment scores in Irish speaking monolinguals. Establishing adult benchmarks for this population is necessary because of the rapid language change of the Irish language which is evident for each new generation of speakers. Furthermore, previous research employing narratives in Irish indicated that when telling a story to their children, adults used some morphosyntactic forms consistently and accurately while other forms they used either inconsistently or inaccurately. Crucially, the forms that parents used consistently and accurately were those that children acquired fully at an early age and used when retelling the same story (Müller et al., 2019; Antonijevic et al., 2020). Lead by those findings, we think that the first step towards creating children's norms for the MAIN in Irish and English is describing the story structure produced by adult Irish-English bilingual speakers, i.e., obtaining benchmarks to which children's narratives will be compared.

The research with multilingual children indicated that the macrostructure scores were similar across their languages (e.g., Hipfner-Boucher et al., 2015; Boerma et al., 2016; Rodina, 2017; Otwinowska et al., 2020). In addition, the macrostructure scores were similar in monolingual adults in different languages (Gagarina et al., 2019). Therefore, we expected the macrostructure scores in both production and comprehension to be similar across the endangered minority language Irish and the dominant language English in adult Irish-English bilinguals. This is the first step toward establishing the developmental trajectory for narrative production and comprehension in Irish-English bilingual speakers.

Materials and methods

The study received full ethical approval from the College of Medicine, Nursing and Health Sciences Research Ethics Committee at the National University of Ireland Galway.

We report here on two studies, both using the MAIN in Irish (O Malley, 2019) and English (Gagarina et al., 2019). The studies used the same procedures and the same participants' inclusion and exclusion criteria and protocols. Study 1 used the Baby Birds and Baby Goats story pair and all participants were teachers in Irish medium education. Study 2 used Cat and Dog story pair and participants were recruited through social media.

Participants

Participants who met the following criteria were invited to participate: healthy adults aged 20–60 years; regular speakers of both Irish and English; the household must have an Internet connection; the household must have a computer or an iPad with a webcam; participants must have or be willing to create a Zoom account; participants must have a quiet space available for the duration of the assessment. Participants could not participate in the study if they had any diagnosis of developmental or acquired language disorder, neurodegenerative, or other conditions that may impact speech, language, or cognitive abilities; or if they spoke daily any other languages in addition to Irish and English. This prevented the potential influence of another language on the participant's narrative while maintaining focus on macrostructural measures of Irish-English multilinguals. Prior to the narrative assessment, all participants completed an online demographic and language questionnaire The Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007) to establish their current language exposure and self-rated language proficiency in Irish and English (refer to Table 1).

Study 1 participants included 14 women and six men, aged between 23 and 54 years ($M = 38.2$, $SD = 14.77$). Participants reported a variation in the current language exposure (refer to Table 1). Three participants reported Irish as their first language (L1) and 17 reported Irish as their second language (L2). Study 2 participants included eight women and two men, aged between 22 and 59 years ($M = 36$, $SD = 12.4$). Irish was L1 of four participants while six participants had Irish as L2. One of the participants involved in the study was a Speech and Language Therapist by profession (TM5), however, they were not familiar with the MAIN. The current language exposure and age of acquisition for Irish as well as the self-rated proficiency for Irish and English are presented in Table 1. Detailed information about language history and proficiency in Irish and English is presented in the Supplementary Materials.

TABLE 1 Demographic and language variables reported through LEAP-Q (Marian et al., 2007) for Study 1 and Study 2.

Study	N	Age	Years in education	Current exposure to Irish	Current exposure to English	Age of Acquisition (AoA) Irish	Age of Acquisition (AoA) English	Proficiency in speaking Irish (0–10)	Proficiency in speaking English (0-10)
1	20	38.2 (14.77)	17.85 (1.76)	38.9 (22.79)	60.85 (22.89)	4.9 (5.47)	0.75 (1.37)	8.75 (0.85)	9.55 (0.83)
2	10	36 (12.4)	18.2 (2.78)	39 (13.05)	60.5 (13)	7 (10.78)	2.5 (2.99)	9.1 (1.29)	8 (1.05)

Participants in Study 1 and Study 2 were matched on age, years of education, current exposure to Irish and English, AoA for Irish and English, and self-rated proficiency in speaking Irish and English (refer to Table 1). Individual scores, means, and SDs for all language related variables reported through LEAP-Q (Marian et al., 2007) are presented in the Supplementary Materials.

Materials and procedure

The MAIN (Gagarina et al., 2019) was the language assessment tool used to collect narrative data. The tool was developed to assess the narrative comprehension and production abilities of bilingual children 3–10 years of age. The MAIN includes four parallel stories (Cat, Dog, Baby Birds, and Baby Goats) each accompanied by a set of six pictures. All four stories include three distinct episodes, where each episode contains five elements: a goal, an attempt, an outcome, and two internal state terms positioned in the sequence as an initiating event and as a reaction. A goal (G) represents a statement of an idea of the protagonists to deal with the initiating event (e.g., "Mother bird wanted to catch worms"). An attempt (A) is an indication of action to obtain the goal (e.g., "Mother bird looked for food"); an Outcome (O) is the event following the attempt and is causally linked to it (e.g., "Mother fed the baby birds"). Internal state terms (ISTs) can be either an initiating event that sets the events of the story in motion (e.g., "Mother bird saw that the baby birds were hungry") or a reaction that defines how the protagonist feels/thinks about the outcome (e.g., "Baby birds were happy/not hungry anymore"). Across the four stories the details related to the protagonists, background and foreground information, and content were controlled to allow for comparison between two languages or between elicitation modes. The MAIN is designed to use one of three elicitation modes: story tell, story retell, and story tell after listening to a different model story (Gagarina et al., 2012, 2019).

The theoretical approach underpinning the MAIN distinguishes two main aspects of macrostructure: story structure and story complexity. Story structure is a quantitative score reflecting the number of episodic elements produced in the narrative. It consists of a score for describing the story setting (a reference to time and place) and scores for elements present in each of the three episodes. Given that each story includes settings referring to time and place that are unique for all three episodes, and the three episodes that each contain G, A, and O as well as IST as an initiating event and IST as a reaction, the story structure score can reach a maximum of 17 (2 for settings and 5 for elements in each of the 3 episodes). While ISTs are included in the MAIN as a part of the macrostructure, they also form a bridge between narrative organization on a more general conceptual level and the linguistic encoding of this information at the lexical level. In addition to ISTs being a part of the story structure score, the MAIN includes a separate ISTs score referring to all instances of perceptual and physiological state terms, consciousness and emotion terms, mental verbs, and verbs of saying and telling.

Narrative comprehension in the MAIN is examined by a set of ten open-ended questions focusing on goals and ISTs, the elements of macrostructure that are not directly present in the pictures but must be inferred (Bohnacker and Gagarina, 2020). Three of the ten questions target G, one from each episode; Six questions target ISTs, three as an initiating event, and three as a reaction. One question focuses on inferencing and requires the participant to reason about the meaning of the whole story (Gagarina et al., 2012). Given that there are 10 questions that each can be awarded one point, the maximum comprehension score is 10 points (Gagarina et al., 2012, 2019).

The narrative assessment was conducted using the MAIN in English (Gagarina et al., 2019) and Irish (O Malley, 2019). The general administration procedure for the MAIN was followed. The order of the languages and the stories across languages were counterbalanced. An online moderated assessment was conducted *via* a professional Zoom account. A custom-made PowerPoint presentation embedding the 6 pictures for each story was used to conduct the narrative procedure and share the pictures with participants (Hamdani et al., 2021). Participants' responses were audio-recorded using the Audacity software. The story tell elicitation method was used. The assessment started with a short warm-up session in the same language as the assessment. After that, participants were shown three envelopes to choose a story. As per the MAIN protocol (Gagarina et al., 2012, 2019), this was done to create an illusion that the researcher did not know the story that the participant was about to tell. In the beginning, they saw all six pictures together to get acquainted with the whole story.

Subsequently, they were asked to tell the story while seeing two pictures (representing one episode of the story) at a time. The researcher remained silent except for the general feedback signals. Following narrative production, participants were asked 10 comprehension questions. The assessment took approximately 15 min per participant. The whole procedure was repeated 1–2 weeks later in the other language. Procedures were identical for Study 1 and Study 2 except that the Baby Birds and Baby Goats story pair was used in Study1 and Cat and Dog story pair in Study 2. All researchers that were involved in data collection had been trained in telehealth administration as a part of their degree in speech and language therapy.

The same researcher conducted the assessment in both languages introducing an aspect of bilingual mode for participants (Grosjean, 1989). This is, however, unavoidable because of the near-universal multilingualism of Irish with English leading to all Irish speakers understanding that their communication partner is not only an Irish but also an English speaker.

Data analyzes

All narratives and answers to comprehension questions were transcribed verbatim. The narratives were then analyzed for the two measures of macrostructure: story structure and ISTs. The answers to the comprehension questions were analyzed separately. Throughout data analysis, researchers referred to the scoring examples in the MAIN: Gaeilge (Irish) (O Malley, 2019) and the MAIN (Gagarina et al., 2019). No points were awarded for the repetition of the same elements.

Identical data analyses were conducted separately for Study 1 and Study 2. The following analyses concerned macrostructure measures: story structure, ISTs, and comprehension score of the MAIN in Irish and English. To address the aim of this study and examine whether there are differences in macrostructure scores across languages a general linear model was conducted with factors: language (Irish/English) and output type (production/comprehension) including story structure and story comprehension scores as dependant variables. Both scores were expressed as proportions, story structure out of 17 and story comprehension out of 10, to allow for direct comparison.

Results

Prior to analyzing data related to the MAIN, participants' language experience obtained by LEAP-Q language questionnaire (Marian et al., 2007) was compared across Irish and English for Study 1 and Study 2. In Study 1, participants' current exposure was higher to English (M = 60.86, SD = 22.89) than to Irish (38.90, SD = 22.79) [$t_{(19)}$ = 2.15, $p = 0.05$, $d = 0.47$] and their self-rated proficiency was also higher in English ($M = 9.55$, $SD = 0.83$) than Irish ($M = 8.75$, $SD = 0.85$) [$t_{(9)} = 2.79$, $p = 0.01$, $d = 0.61$]. AoA for English ($M = 0.75$, $SD = 1.37$) was lower than AoA for Irish ($M = 4.9$, $SD = 5.47$) [$t_{(19)} = -3.237$, $p = 0.004$, $d = -0.724$]. Similarly, in Study 2, participants' current exposure to English ($M = 60.50$, $SD = 13.01$) was higher than to Irish (39.50, $SD = 13.01$) [$t_{(9)} = 2.55$, $p = 0.03$, $d = 0.77$] and their self-rated proficiency was also higher in English ($M = 9.1$, $SD = 1.29$) than Irish ($M = 8$, $SD =$ 1,05) [$t_{(9)} = 2.4$, $p = 0.04$, $d = 0.73$]. However, AoA for English ($M = 2.5$, $SD = 2.99$) was not significantly different from AoA for Irish ($M = 7$, $SD = 10.78$) [$t_{(9)} = -1.12$, $p = 0.293$, $d = -0.35$], which is most likely result of the high variability in AoA for Irish. Language variables reported for Study 1 and Study 2 are presented in Table 1 above.

Comparison of story structure and story comprehension across Irish and English

In Study 1, the mean story structure score in Irish was 11.8 (SD = 2.53) and mean comprehension score in Irish was 9.30 (SD = 1.26); the mean English story structure score was 11.05 (SD = 3.38), and the mean comprehension score in English was of 9.35 (SD = 1.09). To be able to directly compare production and comprehension scores, the raw scores were transformed into proportions out of 17 for story structure and out of 10 for comprehension. A general linear model with factors language (Irish/English) and output type (production/comprehension) indicated no significant difference in the overall performance across languages [$F_{(1,19)} = 0.615$, $p = 0.44$, $\eta^2 = 0.031$]. A significant overall difference was observed for output type [$F_{(1,19)} = 60.85$, $p < 0.001$, $\eta^2 = 0.76$]. Participants performed significantly better in narrative comprehension than production, irrespective of the language. No significant interaction was found between language and output type [$F_{(1,19)} = 1.04$, $p = 0.32$, $\eta^2 = 0.05$] indicating that the discrepancies between production and comprehension scores were the same in both languages (refer to Figure 1).

In Study 2, the overall mean story structure score in Irish was 11.7 (SD = 1.95), and the mean comprehension score in Irish was 9.4 (SD = 0.84); the mean story structure score for English was 12.10 (SD = 1.85), and mean comprehension score for English was 9.4 (SD = 0.52). Similar to Study 1, production and comprehension scores were subsequently transformed into proportions to enable their direct comparison.

A general linear model with factors language (Irish/English) and output type (production/comprehension) indicated no significant difference in the overall performance across languages [$F_{(1,9)} = 0.29$, $p = 0.603$, $\eta^2 = 0.031$]. A significant difference was observed for output type [$F_{(1,19)} = 192.64$, $p <$ 0.001, $\eta^2 = 0.96$]. Participants performed significantly better in narrative comprehension than production, irrespective of

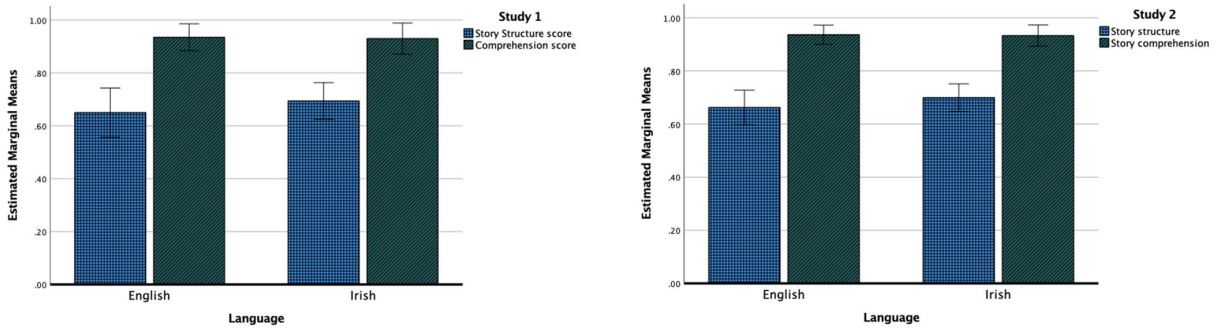

FIGURE 1
The MAIN story structure and story comprehension scores and standard deviations across Irish and English in Study 1 and Study 2.

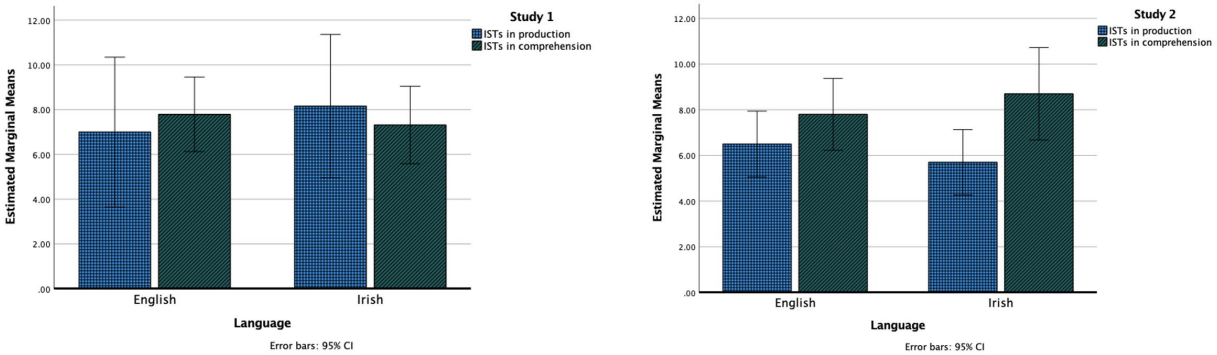

FIGURE 2
The internal state terms (ISTs) in production and comprehension across Irish and English in Study 1 and Study 2.

the language. No significant interaction was found between language and output type [$F_{(1,19)} = 0.159$, $p = 0.70$, $\eta^2 = 0.017$] indicating similar discrepancy between production and comprehension across the languages (refer to Figure 1).

Finally, combined analysis across Study 1 and Study 2 indicated no significant three-way interaction between the language, the type of production, and the story pair [$F_{(2,27)} = 0.339$, $p = 0.715$, $\eta^2 = 0.025$] confirming that similar results were obtained across the two story-pairs Baby Birds and Baby Goats vs. Cat and Dog stories (refer to Figure 1).

Internal state terms across Irish and English

In Study 1 (Baby Birds and Baby Goats story pair), a general linear model with factors language (Irish/English) and output type (production/comprehension) indicated that there was neither significant difference in the number of ISTs across the languages [$F_{(1,19)} = 0.229$, $p = 0.638$, $\eta^2 = 0.013$] nor the number of ISTs produced in comprehension vs. production [$F_{(1,19)} = 0$, $p = 0.938$, $\eta^2 = 0$]. There was no significant interaction between the language and the output type [$F_{(1,19)} = 1.685$, $p = 0.211$, $\eta^2 = 0.086$] (refer to Figure 2).

In Study 2 (Cat and Dog story pair), a general linear model with factors language (Irish/English) and output type (production/comprehension) indicated that there was no significant difference in the number of ISTs across the languages [$F_{(1,9)} = 0.007$, $p = 0.935$, $\eta^2 = 0.001$]. However, a higher number of ISTs was produced in comprehension than in production [$F_{(1,9)} = 6.12$, $p = 0.035$, $\eta^2 = 0.405$]. There was no significant interaction between the language and the output type [$F_{(1,9)} = 0.159$, $p = 0.70$, $\eta^2 = 0.017$] (refer to Figure 2).

Discussion

The aim of the current study was to establish measures of macrostructure in narrative production and comprehension for

Irish-English adult bilinguals and to use those as a baseline for further comparison of narrative macrostructure in Irish-English bilingual children. There were two subsets of data, Study 1 employed the Baby Birds and Baby Goats story pair, and Study 2 employed Cat and Dog story pair from the Multilingual Assessment Instrument for Narratives (MAIN; Gagarina et al., 2012, 2019) to elicit narratives in the tell mode. Despite participants' different backgrounds (participants in Study 1 were teachers in Irish-immersion schools while participants in Study 2 had more diverse linguistic backgrounds), both groups indicated that in the recent past they had higher exposure to English than to Irish and also rated their proficiency in English to be higher than in Irish. This is likely the case due to the convergence of the languages, the universal bilingualism of the Irish language with English, and the global dominance of English. Similar results were previously reported in children in research by Nic Fhlannchadha and Hickey (2021), who concluded that young Irish-English bilinguals from Irish dominant homes are often the minority in Irish immersion education, reflecting that the majority of Irish-English bilinguals have an abundant exposure to English daily. This finding is also in line with the changes in the sociolinguistic landscape of Ireland that have been well documented (Ó Catháin, 2016).

Macrostructure in comprehension and production across Irish and English

As expected on the basis of previous studies including children (e.g., Gagarina et al., 2015; Hipfner-Boucher et al., 2015; Boerma et al., 2016; Bohnacker, 2016; Gagarina, 2016; Rodina, 2017; Otwinowska et al., 2020) and monolingual adults (Gagarina et al., 2019), there was no difference in macrostructure scores across languages in either production or comprehension of narratives. These results were consistent across Study 1 and Study 2. A comparison of narrative macrostructure measures across production and comprehension indicated that overall comprehension scores were higher than production scores, but that this trend did not differ across the languages. In narrative production, mean story structure scores for both English and Irish were in the same range as observed by Gagarina et al. (2019). The mean story structure score was 11.8 for Irish and 11.05 for English concurring with those reported by Gagarina et al. (2019) being in the range of 11–12 points for monolingual German-, Russian-, and Swedish-speaking adults. These findings are encouraging because they suggest that story structure scores are stable across different languages and comparable between monolingual and multilingual adults. Therefore, the obtained scores can be used as a baseline to which children's narrative macrostructure scores will be compared in the future. In that context, it is important to notice that adults did not show the ceiling effect in either production or comprehension and that their narrative comprehension scores were higher than production scores. Similar to the current study, higher comprehension scores relative to production scores were also reported in previous studies (e.g., Bohnacker, 2016; Kapalková et al., 2016). One potential reason for this discrepancy was outlined by Bohnacker (2016) who observed that Goals and ISTs were frequently produced in response to the comprehension questions but rarely spontaneously produced in the narrative production. Goals and ISTs are not explicitly depicted in the elicitation pictures and the narrator needs to infer them from the story plot in narrative production. In narrative comprehension, these elements are specifically addressed in the questions and, therefore, attention is pointed toward them potentially making them easier to include in the answer. Closer inspection of the results of previous studies indicated that this gap between comprehension and production contracted with age (Bohnacker, 2016) and an increase in language exposure (e.g., Roch et al., 2016). In addition, higher story structure scores with more story grammar elements were observed in retell then tell mode (e.g., Kapalková et al., 2016; Roch et al., 2016), which could be a consequence of children hearing explicitly the elements that in the tell mode they would need to infer themselves. The ability to infer the elements not directly present in the pictures is related to the theory of mind and also understanding of the plot of the whole story (Gagarina, 2016). In this respect, both narrative production and comprehension require cognitive in addition to linguistics abilities. While cognitive and linguistic abilities are developing in children, and potentially leading to a reduction in the gap between story structure and story comprehension scores, it is important to know that adult Irish-English bilinguals still achieved higher scores in narrative comprehension than production. This data is an important benchmarking point for the comparison of narrative production and comprehension in Irish-English bilingual children. Narrative macrostructure has been shown to successfully differentiate between TD and DLD monolingual and multilingual children without disadvantaging multilingual TD children (Boerma et al., 2016), which is most likely due to its reliance on cognitive functions such as attention and the theory of mind (Blom and Boerma, 2016; Gagarina, 2016). Taking these findings together with the linguistic, cultural, and socioeconomic variability of Irish speakers leads us to believe that narrative macrostructure is the most optimal tool for language assessment of Irish-English bilingual children.

Finally, the fact that no significant difference in macrostructure measures across the languages and the types of output (comprehension vs. production) were observed for both Baby Birds and Baby Goats, as well as Cat and Dog story pairs, supports the original idea that the MAIN stories were created to have parallel macrostructure with the same number of episodes

and the same episode structure (Gagarina et al., 2012, 2015).

Internal state terms in production and comprehension across Irish and English

Similar to story structure and story comprehension scores, ISTs in narrative production and comprehension were compared across Irish and English. In Study 1 (Baby Birds and Baby Goats stories) similar number of ISTs was observed in production and comprehension, and also across the languages. However, in Study 2 (Cat and Dog stories) a higher number of ISTs was observed in comprehension relative to production and this was the case in Irish and English. While this could indicate differences in the story pairs with respect to the elicitation of ISTs, it is important to note that Study 2 had a smaller number of participants whose backgrounds differed from that of the participants in Study 1. While participants in Study 1 were recruited through the Irish-medium schools in English dominant areas where they worked as teachers, participants in Study 2 were recruited through social media and had more diverse backgrounds so the discrepancy in the number of ISTs in production and comprehension could be driven by participants' characteristics. A higher number of ISTs in narrative comprehension relative to production has been observed by Bohnacker (2016) in Swedish-English bilingual children aged 5–7 years. In the study by Bohnacker, ISTs as initiating events and ISTs as reactions were produced in the majority of cases in comprehension questions, however, they were rarely spontaneously produced in narrative production. Furthermore, the number of ISTs as initiating events increased in the narrative production from age 5 to 7 years, but this was not true for the number of ISTs as reactions (Bohnacker, 2016). ISTs as reactions involve understanding the complete story plot, referring to the theory of mind, and inferring how the characters in the story might feel. Therefore, the findings observed in the current study could be pointing toward the difficulty to infer characters' mental states and including them in the narrative production. A similar type of difficulty was observed in a study that examined another minority language, Gaelic, with respect to inference in reading comprehension. Dickson et al. (2021) found that primary school children in Gaelic-medium education who had English as their dominant language struggled to answer questions requiring them to infer information from a paragraph they read. This difficulty, however, was not observed for both languages of the English-Gaelic bilinguals, which is different from the current study. Discussing the nature of ISTs, Gagarina (2016) suggested that ISTs are much more dependent on lexical knowledge than other macrostructure components. Therefore, the difference in the number of ISTs in production and comprehension observed in the current study could be a result of the potential discrepancy between receptive and expressive vocabulary in both languages for this group of participants. To understand whether the observed pattern of results reflects differences in the two sets of stories or whether it is a result of the characteristics of the participants, all four stories would need to be compared across the same group of Irish-English bilinguals, which will be the aim of future studies.

Code-switching

Narrative assessment is particularly suited for multilinguals because it allows for observation of the phenomena specific to language production in multilinguals such as code-switching (Gagarina et al., 2015). Frequent code-switching with English is a significant characteristic of modern Irish (Ó Catháin, 2016). Code-switching was evident during Irish narrative production and comprehension in the current study. English words were frequently used while beginning the Irish narrative production, and included "OK, so," "OK," and "So." Despite "OK" having a direct Irish translation, "*ceart go leor.*" "OK, so" and "so" do not have direct Irish translations, indicating a lexical gap (Ní Laoire, 2016). Participants may have used these words and phrases to emphasize a point (Ní Laoire, 2016), e.g., at the beginning of the episode. Similarly, numerous participants used code-switching to English to emphasize the end of a sentence or episode, using phrases such as "*Sin alright?*" or "Is that alright?," "*Sin é really*" or "That's it really" and "Em yeah." This type of code-switching indicates metalinguistic awareness where a different language is used to emphasize the change in topic. Some participants used code-switching when they were unaware of the correct Irish expression e.g., using "nest" instead of *nead* and "the cat ran away" instead of *rith an cat leis*. However, due to the bilingual approach of the MAIN Gaeilge (Irish) (O'Malley and Antonijevic, 2020), these responses were marked as correct. Interestingly, participants used code-switching as a form of linking sentences throughout narrative productions, despite those words existing in Irish. Those conjunctions included "yeah," "you know," "because," "and then," "really," "either" and "alright," "is it?" and "is that what you mean?." We suspect that this is likely a result of the almost universal bilingualism of the Irish sociolinguistic context' (O'Malley and Antonijevic, 2020, p. 127), and participants being in a bilingual mode (Grosjean, 1989) as well as knowing that the researcher is also multilingual and will understand their responses in both Irish and English. One participant (CK4) also used the verb "scalaíonn" when describing the cat climbing up the tree. As this is not a verb in Irish, this may be an example of *Béarlachas* or "Englishism," which describes the contact between Irish and English (Ní Laoire, 2016, p. 101). The features of code-switching outlined above are aligned with those described by Ní Laoire (2016), and reports by Ó Catháin (2016) who described the younger generations

of Irish speakers using Irish differently from that of the previous generations.

Online administration of the MAIN

All three researchers who participated in data collection had previous experience with telehealth. They found the administration of the MAIN online to be straightforward. It was helped by the clear instructions, a slideshow of pictorial stimuli, and clear visuals for comprehension questions (Hamdani et al., 2021). The online platform allowed for good rapport building at the beginning of the assessment. Online administration improved time management and allowed for flexibility in arranging assessment times and dealing with cancellations. However, because the home environment being more intimate than a research lab or a clinic, the researchers had the impression that some participants were more guarded. Particularly relevant for Irish settings was that conducting the assessment online allowed for greater geographical reach in recruiting participants. The highest density of Irish speakers is in rural and sometimes remote areas of Ireland. The in-person assessment would hinder their participation because either researchers or participants would have to travel to the place of assessment and in this way significantly increase the time and the cost involved. On the other hand, during in-person assessment researcher has more control over the environment with no risk of internet connection breakdown, difficulties with sound, or participants not having a quiet place in their home. With respect to communication, it was sometimes difficult to read facial expressions or body language. Communication was also sometimes impaired by poor internet connection which could cause overlap when giving instructions and asking comprehension questions. Future studies could work on minimizing the downsides of online assessment given that this mode of assessment has good potential to be used in clinical settings (Peña and Sutherland, 2022).

Limitations

This study has several limitations. The smaller sample size in Study 2 is a limitation and future research should include a larger number of participants. A further limitation is that the MAIN was originally created for an in-person assessment. As a result of the COVID-19 pandemic, the MAIN was subsequently adapted for online use (Hamdani et al., 2021). The online administration of the MAIN was employed in the study by Pratt et al. (2022). They noted during the data collection that participants did not always clearly see what was going on in the pictures and that the size of the pictures needed to be increased. This current study used the original PowerPoint slides (Hamdani et al., 2021) and the picture size was not increased. This problem became clear because on re-examination of the picture stimuli during the comprehension questions, the story was clarified for some participants. Multiple individuals incorrectly described the first episode of the Baby Goats story as the goat "cooling down" or "swimming." However, it became clear to the same participants during the comprehension task that the baby goat was in fact drowning, with one participant (CK8) exclaiming "oh actually, she might be looking for ah help." Not seeing the pictures clearly may have impacted the participants' ability to produce full GAO sequences, and may have resulted in higher mean comprehension scores in comparison to the mean story structure scores. The size of the pictures for online administration should be adjusted in the future to enable participants to better view the details.

Conclusion

The current study aimed to compare macrostructure measures in the MAIN stories (Gagarina et al., 2012, 2019) across output type (production vs. comprehension) in the two languages of Irish-English multilinguals. This study also aimed to establish a baseline for macrostructural measures in Irish and English using the MAIN and MAIN Gaeilge (Irish) (O Malley, 2019) that can be used in future research as well as in clinical settings.

The similarity in macrostructure measures that were obtained across languages during production and comprehension of narratives indicated that differences in language exposure, AoA, and self-rated proficiency between Irish and English did not influence measures of macrostructure in either production or comprehension. Therefore, the results of the current study suggest that the MAIN macrostucture measures are not sensitive to linguistic variability, which is the characteristic of Irish-English multilinguals throughout Ireland and is due to the ever-increasing use of the majority language English and the decreasing use of the minority language Irish. This implies that the MAIN is an optimal language assessment tool for Irish-English bilingual children. The mean story structure and story comprehension scores observed in this study may be cautiously used as a baseline for measures of macrostructure among Irish-English multilinguals. Future studies should focus on using the MAIN (Gagarina et al., 2012, 2019) and MAIN Gaeilge (Irish) (O Malley, 2019) to assess Irish-English bilingual children and determine the developmental trajectories for measures of the macrostructure. The final aim is to provide a valuable tool for language assessment of Irish-English multilingual children in clinical settings, the tool that can overcome challenges of language assessment of a fast changing, endangered minority language Irish. In addition, having an option of online assessment would enable clinicians to reach children across Ireland that are in need of language assessment. The current study is one of the first steps toward that goal.

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 Research Ethics Committee of the College of Medicine, Nursing and Health Sciences, National University of Ireland Galway. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SA designed the research, contributed to the analyses, and writing of the manuscript. SC, CK, and TN contributed to the data collection, data analyses, and writing of the manuscript. All authors contributed to the article and approved the submitted version.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.916214/full#supplementary-material

References

Antonijevic, S., Durham, R., and NÍ Chonghaile, I. (2017). Language performance of sequential bilinguals on an Irish and English sentence repetition task. *Linguist. Approach. Bilingual.* 7, 359–393. doi: 10.1075/lab.15026.ant

Antonijevic, S., Muckley, S. A., and Müller, N. (2020). The role of consistency in use of morphosyntactic forms in child-directed speech in the acquisition of Irish, a minority language undergoing rapid language change. *J. Child Lang.* 47, 267–288. doi: 10.1017/S0305000919000734

Armon-Lotem, S., de Jong, J., and Meir, N. (2015). *Assessing Multilingual Children: Disentangling Bilingualism From Language Impairment*. Bristol: Multilingual Matters.

Berman, R. A., and Slobin, D. I. (1994). *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Mahwah, NJ: Erlbum.

Blom, E., and Boerma, T. (2016). Why do children with language impairment have difficulties with narrative macrostructure? *Res. Develop. Disabil.* 55, 301–311. doi: 10.1016/j.ridd.2016.05.001

Boerma, T., Leseman, P., Timmermeister, M., Wijnen, F., and Blom, E. (2016). Narrative abilities of monolingual and bilingual children with and without language impairment: implications for clinical practice. *Int. J. Commun. Disord.* 51, 626–638. doi: 10.1111/1460-6984.12234

Bohnacker, U. (2016). Tell me a story in English or Swedish: narrative production and comprehension in bilingual preschoolers and first graders. *Appl. Psycholinguist.* 37, 19–48. doi: 10.1017/S0142716415000405

Bohnacker, U., and Gagarina, N. (2020). "Introduction to MAIN–Revised, how to use the instrument and adapt it to further languages" *ZAS papers in Linguistics* (Berlin: ZAS).

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Rev. Educ. Res.* 49, 222–251. doi: 10.3102/00346543049002222

Curenton, S. M., and Justice, L. M. (2004). African American and Caucasian preschoolers' use of decontextualized language: Literate language features in oral narratives. *Lang. Speech Hear. Serv. Sch.* 35, 240–253. doi: 10.1044/0161-1461(2004/023)

Dickson, E., Manderson, L., Obregon, M., and Garraffa, M.. (2021). Tracking biliteracy skills in students attending gaelic medium education: effects of learning experience on overall reading skills. *Languages* 6, 55. doi: 10.3390/languages6010055

Friend, M., and Bates, R. P. (2014). The union of narrative and executive function: different but complementary. *Front. Psychol.* 5, 469. doi: 10.3389/fpsyg.2014.00469

Gagarina, N. (2016). Narratives of Russian-German preschool and primary school bilinguals: Rasskaz and Erzaehlung. *Appl. Psycholinguist.* 37, 91–122. doi: 10.1017/S0142716415000430

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiunien,e, I., et al. (2012). "MAIN: multilingual assessment instrument for narratives," in *ZAS papers in Linguistics* (Berlin: ZAS).

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Bohnacker, U., et al. (2019). "MAIN: multilingual assessment instrument for narratives" in *Revised Version. ZAS Papers in Linguistics* (Berlin: ZAS), 62.

Gagarina, N., Klop, D., Kunnari, S., Tantele, K. V., Bohnacker, U., and Walters, J. (2015). "Assessment of narrative abilities in bilingual children," in *Assessing Multilingual Children Disentangling Bilingualism from Language Impairment, eds* S. Armon-Lotem, J. de Jong, and N. Meir (Bristol: Multilingual Matters), 243–277.

Gagraina, N., Bohnacker, U., and Lindgren, J. (2019). Macrostructural organization of adults' oral narrative texts. *ZAS Papers Linguist.* 62, 190–208. doi: 10.21248/zaspil.62.2019.449

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain Lang.* 36, 3–15.

Hamdani, S., Kan, R., Chan, A., and Gagarina, N. (2021). *Online Moderated Assessment Using PowerPoint Presentation*. [Unpublished].

Heilmann, J., Miller, J. F., Nockerts, A., and Dunaway, C. (2010). Properties of the narrative scoring scheme using narrative retells in young school-age children. *Am. J. Speech Lang. Pathol.* 19, 154–166. doi: 10.1044/1058-0360(2009/08-0024)

Hipfner-Boucher, K., Milburn, T., Weitzman, E., Greenberg, J., Pelletier, J., and Girolametto, L. (2015). Narrative abilities in subgroups of English language learners and monolingual peers. *Int. J. Bilingual.* 19, 667–692. doi: 10.1177/1367006914534330

Kapalková, S., Polisensk,á, K., Markov,á, L., and Fenton, J. (2016). Narrative abilities in early successive bilingual Slovak-English children: a cross-language comparison. *Appl. Psycholinguist.* 37, 145–164. doi: 10.1017/S0142716415000454

Lindgren J. (2019). Comprehension and production of narrative macrostructure in Swedish: A longitudinal study from age 4 to 7. *First Lang.* 39, 412–432. doi: 10.1177/0142723719844089

MacWhinney, B. (2005). "A unified model of language acquisition," in *Handbook of Bilingualism*, eds J. F. Kroll and A. M. B. De Groot (Oxford: Oxford University Press, 49–67.

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). the language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)

Müller, N., Muckley, S. A., and Antonijevic-Elliott, S. (2019). Where phonology meets morphology in the context of rapid language change and universal bilingualism: Irish initial mutations in child language. *Clin. Linguist. Phonet.* 33, 3–19. doi: 10.1080/02699206.2018.1542742

Nic Fhlannchadha, S., and Hickey, T. M. (2019). Assessing children's proficiency in a minority language: exploring the relationship between home language exposure, test performance and teacher and parent ratings of school-age Irish-English bilinguals. *Lang. Educ.* 33, 1–20. doi: 10.1080/09500782.2018.1523922

Nic Fhlannchadha, S., and Hickey, T. M. (2021). Where are the goalposts? Generational change in the use of grammatical gender in Irish. *Languages* 6, 1–23. doi: 10.3390/languages6010033

Nippold, M. A., Ward-Lonergan, J. M., and Fanning, J. L. (2005). Persuasive writing in children, adolescents, and adults: a study of syntactic, semantic and pragmatic development. *Lang. Speech Hear. Serv. Sch.* 36, 125–138. doi: 10.1044/0161-1461(2005/012)

Ní Laoire, S. (2016). "Irish-English code-switching: a sociolinguistic perspective," in *Sociolinguistics in Ireland,* ed R. Hickey (London, UK: Palgrave Macmillan), 81–106.

Ó Catháin, B. (2016). "The Irish language in present-day Ireland," in *Sociolinguistics in Ireland,* ed R. Hickey (London, UK: Palgrave Macmillan), 41–59.

O Malley, M. P., NÍ Choirbín, A., Antonijevic, S., Brennan, S., Mullin, M., Ó Loideáin, C., et al. (2019). "MAIN: Gaeilge (Irish)," in *MAIN: Multilingual Assessment Instrument for Narratives - Revised. Materials for use. ZAS Papers in Linguistics*, eds N. Gagarina, D. Klop, S. Kunnari, K. Tantele, T. Välimaa, U. Bohnacker, and J. Walters. doi: 10.21248/zaspil.56.2019.414

O'Malley, M.-P., and Antonijevic, S. (2020). Adapting MAIN to Irish (Gaeilge). *ZAS Papers Linguist.* 64, 127–138. doi: 10.21248/zaspil.64.2020.565

Otwinowska, A., Mieszkowska, K., Biaecka-Pikul, M., Opacki, M., and Haman, E. (2020). Retelling a model story improves the narratives of Polish-English bilingual children. *Int. J. Bilingual. Educ. Bilingual.* 23, 1083–1107. doi: 10.1080/13670050.2018.1434124

Paradis, J., Genesee, F., and Crago, M. (2010). *Dual Language Development and Disorders: A Handbook on Bilingualism and Second Language Learning*. Baltimore, MD: Brookes.

Peña, E., Gillam, R., and Bedore, L. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *J. Speech Lang. Hear. Res.* 57, 2208–2220. doi: 10.1044/2014_JSLHR-L-13-0151

Peña, E., and Sutherland, R. (2022). Can you see my screen? Virtual assessment in speech and language. *Lang. Speech Hear. Serv. Sch.* 53, 329–334. doi: 10.1044/2022_LSHSS-22-00007

Pesco, D., and Kay-Raining Bird, E. (2016). Perspectives on bilingual children's narratives elicted with the Multilingual Assessment Instrument for Narratives. *Appl. Psycholinguist.* 37, 1–9. doi: 10.1017/S0142716415000387

Pratt, A. S., Anaya, J.B., Ramos, M. N., Pham, G., Muñoz, M., Bedore, L. M., et al. (2022). From a distance: comparison of in-person and virtual assessments with adult-child dyads from linguistically diverse backgrounds. *Lang. Speech Hear. Serv. Sch.* 53, 360–375. doi: 10.1044/2021_LSHSS-21-00070

Roch, M., Florit, E., and Levorato, C. (2016). Narrative competence of Italian-English bilingual children between 5 and 7 years. *Appl. Psycholinguist.* 37, 49–67. doi: 10.1017/S0142716415000417

Rodina, Y. (2017). Narrative abilities of pre-school bilingual Norwegian-Russian children. *Int. J. Bilingual.* 21, 617–635. doi: 10.1177/1367006916643528

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 29 September 2022
DOI 10.3389/fpsyg.2022.897031

OPEN ACCESS

EDITED BY
Natalia Gagarina,
Leibniz Center for General Linguistics (ZAS), Germany

REVIEWED BY
Jidong Chen,
California State University, Fresno, United States
Nan Xu Rattanasone,
Macquarie University, Australia
Li Sheng,
Hong Kong Polytechnic University, Hong Kong SAR, China

*CORRESPONDENCE
Jiangling Zhou
jiangling.zhou@cuhk.edu.hk

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

RECEIVED 15 March 2022
ACCEPTED 22 August 2022
PUBLISHED 29 September 2022

CITATION
Zhou J, Mai Z, Cai Q, Liang Y and Yip V (2022) Reference production in Mandarin–English bilingual preschoolers: Linguistic, input, and cognitive factors. *Front. Psychol.* 13:897031. doi: 10.3389/fpsyg.2022.897031

Reference production in Mandarin–English bilingual preschoolers: Linguistic, input, and cognitive factors

Jiangling Zhou[1]*, Ziyin Mai[2], Qiuyun Cai[2], Yuqing Liang[2] and Virginia Yip[1]

[1]Childhood Bilingualism Research Centre, Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China, [2]Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, Hong Kong SAR, China

Reference in extended discourse is vulnerable to delayed acquisition in early childhood. Although recent research has increasingly focused on effects of linguistic, input, and cognitive factors on reference production, these studies are limited in number and the results are mixed. The present study provides insight into bilingual reference production by investigating how production of referring expressions in the two languages of preschool bilingual children may be influenced by structural similarities and differences between the languages, frequency of referring expressions in maternal input, amount of exposure to each of the languages, and working memory capacity. Using two stories in the Multilingual Assessment Instrument for Narratives (MAIN), we examined character introduction and re-introduction in oral narratives of 4–6-year-old Singaporean bilingual children acquiring Mandarin Chinese and English (n = 21), and in child-directed speech of the mothers (n = 17). The children's language exposure, executive function, and general bilingual proficiency were also recorded or directly tested through structured interviews with the parents or standardized assessments with the children. Data collection was conducted remotely in real time over a video-conferencing platform, supplemented by on-site audio recording to ensure sound quality. Results showed prolonged development in the production of felicitous REs for first mentions and over-reliance on overt marking of definiteness in our bilingual children. Mixed modeling revealed that frequency of felicitous REs in the input predicted children's production of felicitous REs across languages and discourse functions, with a modulating effect of working memory. Overall, our findings are consistent with previous ones in that reference production is vulnerable in early Mandarin-English bilinguals in a multilingual society. This study also presents novel evidence that structural frequency in the input interacts with working memory in shaping patterns of reference production in bilingual children.

KEYWORDS

bilingual reference production, input, frequency, amount of language exposure, working memory, cross-linguistic influence, Mandarin, English

Introduction

Reference is one of the core aspects of human communication. A variety of linguistic structures such as lexical noun phrases (NPs, e.g., the goat), demonstratives (e.g., this), and personal pronouns (e.g., she) can serve as referring expressions (REs). To produce felicitous REs, speakers must develop sensitivity to language-specific constraints at syntactic, semantic, and discourse-pragmatic levels as well as the cognitive ability of perspective taking.

To introduce a new referent into discourse—for instance, a fox known to the speaker but not shared by the listener, adult speakers prefer indefinite expressions (e.g., *There is a fox hiding behind the tree*), rather than definite ones (e.g., *The fox is hiding behind the tree*). A lengthy period of development has been documented in children before adult-like use of REs in extended discourse (e.g., narrative production), with significant developmental changes occurring after 7–10 years (Hickmann et al., 2015). Monolingual children under 5–6 years have been shown to overuse definite nominals. They produced a substantial number of NPs with a definite determiner in article languages like English (e.g., Hickmann et al., 1996) or used inappropriate bare nouns interpretable as definite in article-less languages like Mandarin (e.g., Min, 1994; Wu et al., 2015) in contexts where the intended referents were new and unknown to the listener. Apart from choice of elements inside the NP, adult speakers manipulate word order to mark the new/old distinction, preferring the "old-before-new" word order. Young children, however, exhibit a preference for ordering new information before old information (e.g., German: Narasimhan and Dimroth, 2008) or display no ordering preference (e.g., English: Chen and Narasimhan, 2018; Mandarin: Chen et al., 2020). Though not always the case, bilingual children have been reported to show uses of REs that differentiate them from monolingual children, producing non-target-like forms unattested in monolingual children (Zhou et al., 2021; Zhou and Yip, 2021), and using linguistic forms to overtly mark definiteness to an excessive extent (Aalberse et al., 2017).

Different attempts have been made to account for children's late mastery of reference in extended discourse situations and differences among individuals, examining the role of cross-linguistic influence, language input, and cognitive capacities. Although several recent studies on reference production have investigated the effects of input and/or cognitive factors (e.g., Jia and Paradis, 2015; Lindgren et al., 2020; Serratrice and De Cat, 2020), these studies are limited in number and the results are mixed. Research adopting a multifactorial perspective to reference production is in its infancy. We attempt to bring different lines of research on reference development closer together by investigating reference production in bilingual children, considering influence of linguistic, input, and cognitive factors. Specifically, we studied bilingual preschoolers speaking Mandarin Chinese (hereafter Mandarin) and English as well as their parents in Singapore, a multilingual society where English and Mandarin are widely spoken. 74.3% of the resident population in Singapore are Chinese, most of whom use English (47.6%) or Mandarin (40.2%) as the most frequently spoken language at home, and 75.5% of the Chinese who speak English most frequently at home also use Mandarin as the second most frequently spoken language at home (Singapore Department of Statistics, 2020). Unlike English, a non-pro-drop language with dedicated morphology to express definiteness, Mandarin encodes definiteness *via* word order, discourse context, and optional use of functional items (e.g., demonstratives). The central question in this article is how linguistic, input, and cognitive factors interact and shape reference production in Mandarin-English bilingual preschoolers in a multilingual society.

Another innovative feature of this study is that we elicited child and adult discourse remotely over an audio-video platform in real time (online) in lieu of the traditional face-to-face methods, which were rendered less feasible, if at all possible, due to social distancing during the COVID-19 pandemic. Remote online assessment has been shown to yield results comparable to face-to-face methods in tests of intellectual abilities, vocabulary and comprehension with preschool and school-age children (e.g., Hodge et al., 2019; Kronenberger et al., 2021; Werfel et al., 2021). However, little has been reported on the feasibility and validity of eliciting narratives from young children through virtual meetings in real time. This study will provide valuable data for future comparison of referential strategies used by children between face-to-face and videoconference-based modalities.

In the following, we review studies on the linguistic, input, and cognitive factors involved in reference production respectively, and present the research goals and methods of the current study, followed by results and discussions on the relation between reference production and the three sets of factors in each of the target languages.

Reference production: Linguistic, input, and cognitive factors

Referential choice in Mandarin and English: Form, function, and acquisition

A speaker's referential choice usually reflects their assumptions about the extent to which a referent is linguistically retrievable or cognitively accessible to the addressee (Ariel, 1990; Gundel et al., 1993). Referents that are deemed more accessible (e.g., receiving shared visual focus of attention, made prominent by preceding discourse environment, and bearing the thematic role of agent) are likely pronominalized, while referents with low accessibility tend to be denoted by nominals (Allen et al., 2008).

In a narrative context, referent accessibility is often discussed in association with discourse function; that is, whether the RE mentions a referent for the very first time (referent introduction/INTRO), maintains reference to an already mentioned referent (reference maintenance), or re-mentions a referent after focusing on a different referent in intervening utterances (re-introduction/

Re-INTRO) (Hickmann et al., 2015). This study will focus on INTRO and Re-INTRO, both of which involve reference to an entity that is outside the addressee's current focus of attention and have been found to pose greater challenges for children than reference maintenance (Wong and Johnston, 2004; Chen and Lei, 2012; Colozzo and Whitely, 2014).

Regarding referential forms, pronominals (i.e., demonstratives and personal pronouns), and null forms[1] neither signal new information nor fulfill the dual purposes of signaling given information while acknowledging a topic shift (Colozzo and Whitely, 2014). They are more suitable for maintenance of reference, and less preferable than nominals for either INTRO or Re-INTRO. Definite nominals presuppose the listener's knowledge whereas indefinite ones do not. Given this, it is natural that indefinite nominals are preferred in INTRO contexts to introduce new referents and definite nominals in Re-INTRO contexts to shift the topic and bring forward previously mentioned referents.

In both Mandarin and English, there are identifiable nominals which have interpretable reference (definite or specific) independent of syntactic position, such as demonstrative NPs (e.g., *zhe zhi yang* "this goat"), kinship terms (e.g., *mama* "mom"), and complex NPs containing a possessor [e.g., *ta (de) mama* "her mother"], a relative clause (e.g., *zai chi cao de yang* "the goat that is eating grass"), or adjectival modification (e.g., *niaochao li de xiaoniao* "the birds in the nest"). An interesting fact about Mandarin demonstrative NPs is that the distal demonstrative *na* "that" is arguably going through a grammaticalization process, in which it has developed additional functions that are typically served by definite articles in languages like English (Chen, 2004). For instance, unlike the demonstrative *na* in (1a), which expresses a distal meaning, *na* in (1b) is deictically neutral, serving as a determiner of a complex NP. This is also reflected by different translations in English in (1a) and (1b), where the demonstrative *na* is felicitously translated into *that* and *the, respectively.* Mandarin demonstrative NPs sometimes appear without the noun in the form of [demonstrative-classifier], functioning like a demonstrative pronoun as in (1c).

1a. Demonstrative NP used deictically.

> Zhe/Na zhi yang hen ke'ai.
> this/that CL goat very cute
> "This/That goat is very cute."
> (Context: the speaker refers to a goat nearby/from a distance.)

1b. Demonstrative NP in deictically neutral contexts.

> Yang mama faxian le na zhi duo zai shu houmian de huli.
> goat mother discover LE that CL hide at tree back DE fox
> "Mommy goat saw the fox that was hiding behind the tree."

1c. [demonstrative-classifier] functioning like a demonstrative pronoun.

> Zhe/Na ge shi shenme?
> this/that CL is what
> "What is this/that?"

While there are structures in which Mandarin and English overlap in both form and function, there exist language-specific structures for expressing definiteness. In English, definiteness is marked by definite/indefinite/numeral determiners distinguishing given from new referents [(Def./Indef./Num. determiner-NP); e.g., *the fox*, *a fox*, *and two foxes*]. Mandarin is an article-less language which does not have such a mechanism for denoting definiteness. Instead, definiteness marking is achieved through a number of nominal structures and their positioning in relation to their subcategorizing verbs. Regardless of RE type, new information typically appears postverbally in Mandarin. Below we will present two such structures [(bare noun) and (numeral-classifier-noun)] and show how their referential meaning changes when they appear pre- and postverbally.

Unlike in English, bare nouns are allowed in Mandarin and interpretable as definite or indefinite, depending on whether they are preverbal or postverbal (Cheng and Sybesma, 1999).[2] Preverbal bare nouns tend to be definite, as shown by *huli* "the fox" in (2), whereas postverbal bare nouns such as *shanyang* "goat" in (2) tend to receive an indefinite reading, unless when they refer to already mentioned or known referents (e.g., in Re-INTRO contexts).

2. Bare noun used for INTRO contexts (definite preverbally, indefinite postverbally)

> Huli xiang chi shanyang.
> fox want eat goat
> "The fox wanted to eat a goat."

A second structure encoding (in)definiteness in Mandarin but not in English is the [Num-Cl-N] structure consisting of a numeral (Num), a classifier (Cl), and a noun (N), as shown by *yi zhi huli* "a fox" in (3). Like bare nouns, postverbal [Num-Cl-N] structures are interpreted as indefinite, unless in contexts where the intended referent is already mentioned and identified. Preverbal [Num-Cl-N] structures usually receive a definite interpretation, as shown by *liang zhi xiaoniao* "the two little birds" in (4a). The [Num-Cl-N] is indefinite when the numeral is *yi* "one"

1 Mandarin allows the use of null forms for referents that are readily identifiable in the immediately preceding discourse or physical context. However, seemingly appropriate null forms in Re-INTRO contexts does not necessarily reflect good discourse integration ability in the speaker because the appropriateness of null forms may reflect the minds of the listener more than the minds of the speaker.

2 If a bare noun follows a verb denoting unbounded states, e.g., *xihuan* 'like', it receives a generic reading (Sybesma, 1992).

(i.e., [*yi*-Cl-N]) and the [*yi*-Cl-N] cannot be placed preverbally or used for Re-INTRO as in (4b).

3. Postverbal indefinite [Num-Cl-N] (numeral optional when it is *yi* "one")

> Yang mama kanjian (yi) zhi huli.
> goat mother see one CL fox
> "Mommy goat saw a fox."

4. Preverbal definite [Num-Cl-N] (impossible when the numeral is *yi*)

> a. Liang zhi xiaoniao kanjian xiaomao lai le, hen haipa.
> two CL little bird see little cat come LE very scared
> "Seeing the little cat coming, the two little birds were very scared."
>
> b. *Yi zhi xiaoniao kanjian xiaomao lai le, hen haipa.
> one CL little bird see little cat come LE very scared
> Intended: "Seeing the little cat coming, the little bird was very scared."

In addition to appearing in canonical Subject-Verb-Object (SVO) sentences as in (2–3), both bare nouns and [Num-Cl-N] structures characteristically occupy postverbal positions in existential *you*-sentences and Subject-Verb (SV) inversion sentences to introduce new referents (Li and Thompson, 1981). This is consistent with their indefinite interpretation in the postverbal position, illustrated in (5) and (6), where *(yi zhi) huli* "(a) fox" appears after the existential verb *you* "have" and the motion verb *lai* "come" respectively. A summary of the nominal expressions used for INTRO and Re-INTRO contexts in Mandarin and English is given in Table 1.

5. Bare noun and [Num-Cl-N] in existential you-sentence.

> You (yi zhi) huli duo zai shu houmian.
> have one CL fox hide at tree back
> "There is a fox hiding behind the tree."

6. Bare noun and [Num-Cl-N] in SV inversion.

> Lai le (yi zhi) huli.
> come LE one CL fox
> "A fox came."

The aforementioned differences in reference coding have been shown in adult Chinese/English speakers' narrative production. In Hickmann et al. (1996), the Chinese speakers mostly used postverbal [Num-Cl-N] [e.g., (3, 5–6)] for INTRO, while the English speakers marked most of the INTROs with an indefinite determiner (likely postverbal, but less frequently compared to Chinese speakers). Hickmann and Hendriks (1999) found that most nominals denoting previously mentioned referents in Re-INTRO contexts were bare nouns [e.g., (2)] and demonstrative NPs [e.g., (1a)] in Chinese speakers and [Def. determiner-NP] in English speakers. These findings confirm that definite/indefinite determiners are the primary mechanism for marking the given/new distinction in nominals in English, while a number of morphosyntactic structures collaborate with word order to mark that distinction in Mandarin.

For children, previous studies showed that monolinguals overproduce definite nominals in English/Mandarin (Hickmann et al., 1996; Wu et al., 2015) and differ from adults by showing no preference for the "old-before-new" word order (Chen and Narasimhan, 2018; Chen et al., 2020). For Mandarin-English bilingual children, a question is how they cope with dual input in developing target-like reference use. Cross-linguistic influence, specific language input, and cognitive capacities have featured frequently in recent literature. We will review cross-linguistic influence in the rest of this section and the input and cognitive factors in the next two sections.

TABLE 1 Nominal expressions and their discourse functions in Mandarin and English.

Nominal expression	Discourse function	Linguistic form	Position	Mandarin	English
Indefinite	[INTRO]	Bare noun[a]	Postverbal	+	N/A
		[Num-Cl-N]	Postverbal	+	N/A
		[Indef. determiner-NP]	Pre/postverbal	N/A	+
		[Num. determiner-NP]	Pre/postverbal	N/A	+
Definite/Identifiable	[Re-INTRO]	Bare noun	Pre-verbal	+	N/A
		[Num-Cl-N]	Pre-verbal	+	N/A
		[Def. determiner-NP]	Pre/postverbal	N/A	+
		Demonstrative NP	Pre/postverbal	+	+
		Other nominals with interpretable reference[b]	Pre/postverbal	+	+

"+," allowed; N/A, non-applicable; Shaded cells are structures in which Mandarin and English overlap in terms of form and function.
[a]Bare noun and [Num-Cl-N] for first mentions of referents tend to be interpreted as indefinite postverbally and as definite preverbally. They receive a definite reading when referring to already mentioned referents. Referential [(yi-)Cl-N] is indefinite.
[b]Other nominals with interpretable reference regardless of syntactic position include kinship terms and complex NPs containing a possessor, a relative clause or adjectival modification.

Cross-linguistic influence (CLI) is likely to take place in domains involving syntax-discourse interface when there is structural overlap between two languages being acquired by the bilingual child (Hulk and Müller, 2000). In this light, reference production is predicted to be vulnerable to CLI. Indeed, evidence for CLI has been reported in previous studies on bilingual children learning Mandarin and an article language. For example, Mai et al. (2021) found that heritage Mandarin children (aged 4–14) in the United Kingdom produced significantly more demonstrative NPs in a syntactic position requiring definite or specific NPs. The authors attributed this difference to possible CLI from English, which obligatorily marks definiteness through overt markers. In Aalberse et al. (2017), heritage Mandarin speakers (aged 15–27) in the Netherland also showed a significant increase in the use of demonstrative NPs in oral narratives, compared to homeland speakers. It was suggested that demonstrative pronouns in Mandarin might have been reinterpreted as definite articles by the heritage speakers due to influence of Dutch, which has dedicated morphology to encode definiteness. Both studies point toward CLI from the language with overt definiteness marking (English, Dutch) to Mandarin. Looking beyond Mandarin, the use of demonstratives as an equivalent of definite articles has been found in other article-less languages, such as Russian, Malay, and Polish in contact with article languages (Polinsky, 2006; Moro, 2016; Otwinowska et al., 2020). These findings are invariably consistent with possible influence of an article-language on an article-less language with respect to definiteness marking.

The above studies either investigated older school-age children or included children with a wide age span, and the target language was a minority language mainly spoken at home. It remains open whether Mandarin-English bilingual preschoolers in a multilingual society where both target languages are spoken would exhibit over-reliance on overt marking of definiteness in Mandarin and behaved similarly to monolinguals regarding pre/postverbal positioning for first mentions (i.e., INTROs), which brings us to input-related factors in bilingual referential choice.

Language exposure and caregiver input in bilingual acquisition

Compared to monolingual children, the input available to bilingual children is proportionally less in each language and typically unevenly distributed across the relevant languages. In cases where the linguistic input is presumably provided by caregivers who are non-native speakers or speak a contact variety of the language, it may also differ from the input monolingual children typically receive in that language in terms of quality (Paradis and Navarro, 2003; Fernald, 2006). Even within bilingual children, the input varies both quantitatively and qualitatively (e.g., presence of school-age older siblings in the home, one or both parents are native speakers of the target language), leading to individual differences in the rate of language growth (Hoff et al., 2014).

Accumulating evidence shows an enormous impact of language input on acquisition outcomes across various linguistic domains (Grüter and Paradis, 2014). The role of input on bilingual reference production is however less clear. For bilingual children who spoke different L1s at home and were schooled exclusively in English in the UK (5-7-year-old), those with greater exposure to English were better at providing informative REs than those with less exposure to English (Serratrice and De Cat, 2020). For heritage speakers of Mandarin (6;9–10;10), those who arrived in Canada at an older age and had a richer and more diverse Mandarin environment at home demonstrated superior performance with INTROs in Mandarin (Jia and Paradis, 2015). Nevertheless, these results contrast with Lindgren et al. (2020), who did not find a significant effect of amount of language exposure on Swedish-German bilingual children (4;0–6;11) in their use of indefinite NPs to introduce referents in either language. It was hypothesized that the null effect of language exposure could be due to typological similarities between Swedish and German in the use of REs for INTRO and the children's relatively high proficiency in both languages.

Note that these studies invariably measured input based on retrospective parental report on the amount and source of the input. Parental report is a valid method to document and calculate coarse-grained input variables (Paradis, 2017). However, it inevitably oversimplifies the picture, as what is actually heard by the children is not captured. A fine-grained transcript-based analysis of real-life child-directed speech would enable us to obtain a more precise understanding of the ways in which input influences children's reference production. Few existing research on reference production has adopted a fine-grained approach to input. An exception is Paradis and Navarro (2003), who analyzed spontaneous language data from one Spanish-English bilingual child (1;9–2;6), two Spanish monolingual children (1;8–2;7 and 1;8–1;11) and their parental interlocutors. Among many fine-grained variables of the input, they measured STRUCTURAL FREQUENCY, which has been reported to positively correlate with acquisition of grammatical structures such as *wh*-questions, relative clauses, and passives (see Ambridge et al., 2015, for a review). They found that not only the bilingual child produced more overt subjects than the monolingual children in Spanish, but the parents of the bilingual child also used overt subjects at a higher rate than the parents of the monolingual children. The findings suggested that differential patterns in the bilingual children's referential choice may be influenced by how often relevant structures are provided in the parental input. The potential effect of structural frequency warrants further investigations with a larger sample including children with different language profiles and their caregivers, which is a motivation of our study.

Working memory and reference production

When telling a story, in addition to accessing appropriate lexical and syntactic forms, a speaker must attend to the target

referent, monitor for differences in the addressee's perspective, and integrate visual and verbal information into a coherent situation model; furthermore, they must maintain and update the situation model by retaining information associated with a discourse referent and retrieving and updating this information in subsequent mentions of the referent (De Cat, 2015). This is a complex set of operations requiring attentional resources and support of executive functions—higher-order cognitive skills for planning and executing complex tasks (Pennington and Ozonoff, 1996; Miyake et al., 2000).

In particular, working memory possibly underpins the use of REs by allowing for an interlocutor to store and update the addressee's perspective and check for convergence by comparing it with one's own perspective (Serratrice and De Cat, 2020). The hypothesis is that when the communication task generates excessive cognitive demand for the parser's working memory, they revert to a more "egocentric" mode (Nilsen and Bacso, 2017) and produce inadequate REs. Nevertheless, the findings have been mixed as to the role of working memory in reference use. In monolingual populations, children, adolescents, and adults with weaker working memory capacity have been shown to encounter greater difficulty in perspective-taking (e.g., Lin et al., 2010; Wardlow and Heyman, 2016; Nilsen and Bacso, 2017). Additionally, computational modeling studies have found that a simulated low working memory model would produce significantly more underspecified REs than a high working memory model (van Rij, 2012; Hendriks, 2016). Further evidence of a positive correlation between working memory and reference production comes from Torregrossa (2017), who found that German monolingual children (8-10-year-old) with lower working memory capacity were less adequate in the production of demonstrative pronouns in oral narratives. However, in Nilsen and Graham (2009), working memory was not predictive of English-speaking children's (4-5-year-old) use of disambiguating modifiers when there was shared access to a referential alternative.

Mixed findings have also been reported in studies on bilingual reference production. Serratrice and De Cat (2020) reported that working memory positively correlated with 5-7-year-old bilingual children's ability to use informative REs for anaphoric reference in English in the presence of a discourse competitor. Torregrossa et al. (2021), however, did not observe any correlation between updating skill (which hinges on working memory) and the production of underspecified pronouns (null subjects and clitics) in Greek in an elicited narration task with Greek-Albanian, Greek-English, and Greek-German bilingual children (7-13-year-old).

Possibly the mixed findings were due in part to differences in the experimental design, operationalization of working memory, and/or age of the participants. Nilsen and Graham (2009) and Serratrice and De Cat (2020) studied preschoolers performing referential communication tasks, and measured cognitive skills by memory of objects and/or backward digit span (BDS). Torregrossa (2017) and Torregrossa et al. (2021) elicited oral narratives from school-age children, with cognitive skills measured by BDS and a 2-back task, respectively.

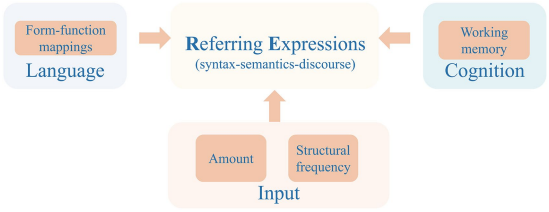

FIGURE 1
Research framework of the current study: Reference use and its relationship with linguistic, input, and cognitive factors.

The interaction between the linguistic and cognitive abilities of the speaker is already particularly intricate (Hendriks, 2016). Such interaction between language and cognition in bilingual children is further complicated by factors such as input and language dominance. In Torregrossa et al. (2021), for example, children who were dominant in Greek produced more overspecified full nouns as a function of lower updating skills, but such effect was absent in children who were dominant in other languages. It was argued that the effects of updating skills were overshadowed by the effects of language exposure in these children, since dominant experience in other languages led to the same pattern of outcomes as lower updating skills in terms of the use of full nouns in Greek—that is, children who were more dominant in other languages showed a stronger tendency of using overspecified full nouns in Greek, regardless of updating skills.

Studies adopting a multifactorial approach, therefore, provide a window into the interplay between linguistic, input, and cognitive factors in bilingual reference production. The findings will shed light on the sources of bilingual-monolingual differences as well as individual differences in reference development. To this aim, we elicited narration from Mandarin-English bilingual preschoolers and collected child-directed speech data by recording storytelling by their mothers.[3] We examined children's production of REs at lexical, syntactic and discourse levels, and investigated its relations with maternal input (in terms of structural frequency), amount of language exposure, and working memory in each of the target languages. Figure 1 illustrates our research framework.

The study

Research questions and predictions

This study investigates 4–6-year-old Singaporean Mandarin-English bilingual children's referential choice for INTRO and Re-INTRO in oral narratives, and examines the contribution of linguistic, input, and cognitive factors to bilingual reference production. Our specific research questions are:

3 Collection of child-directed speech from those children's teachers at the school (i.e., teacher input) is in preparation.

Linguistic factors

What types of REs do Mandarin-English bilingual children use in INTRO and Re-INTRO contexts, respectively, in each target language? Do they show any position (preverbal, postverbal) preference for INTRO?

Predictions

Preferable REs for INTRO and Re-INTRO are indefinite nominals and definite/identifiable nominals, respectively. Considering the persistent overuse of inadequate REs in monolingual preschoolers shown in previous studies (Hickmann et al., 1996; Wu et al., 2015), we expect similarly non-target-like use of REs in our bilingual children—namely, over-production of [Def. determiner-NP] in English and NPs interpretable as definite in Mandarin in INTRO contexts.

Previous research showed over-reliance on overt markers to express definiteness in Mandarin heritage speakers due to cross-linguistic influence of English (e.g., Aalberse et al., 2017; Mai et al., 2021). If influence of English also occurs in our bilingual children, they will produce a high frequency of demonstrative NPs as older heritage Mandarin speakers in previous studies did.

English/Mandarin monolingual children were less likely to use the "old-before-new" word order than adults (Chen and Narasimhan, 2018; Chen et al., 2020). Given this, we expect no preference for postverbal INTROs over preverbal INTROs in either language of the bilingual children.

Input factors

How does Mandarin-English bilingual children's reference production compare to the maternal input? Do they correlate in terms of structural frequency of REs? To what extent is Mandarin-English bilingual children's referential choice influenced by the amount of exposure they receive in each target language?

Predictions

Mother–child differences are expected since the children are predicted to overproduce definite nominals for INTRO, show excessive use of overt markers to express definiteness, and display no preference for postverbal INTROs.

Considering the frequency effect of input observed in Paradis and Navarro (2003), we expect that the structural frequency of REs in maternal input will be reflected in bilingual children's production.

We expect that amount of language exposure predicts bilingual children's production of indefinite nominals for INTRO and definite/identifiable nominals for Re-INTRO in each language, given that previous findings showed a significant effect of amount of exposure in reference production in heritage speakers of Mandarin (Jia and Paradis, 2015) and in bilingual children acquiring English as an additional language (Serratrice and De Cat, 2020).

Cognitive factor

To what extent is Mandarin-English bilingual children's reference production in each of the target languages influenced by working memory?

Predictions

Given the evidence that children with stronger working memory capacity are better able to produce felicitous REs (e.g., Torregrossa, 2017; Serratrice and De Cat, 2020), we expect that bilingual children's working memory capacity predicts their production of indefinite nominals for INTRO and definite/identifiable nominals for Re-INTRO in Mandarin and in English.

Participants

We recruited Mandarin-English bilingual children from a kindergarten in Singapore, where they were enrolled in a full-day Mandarin-English bilingual program, with roughly equal distribution of exposure to each language at school. Their class teachers were native speakers of either Mandarin or English and were assigned to address the children in their native language. A screening questionnaire was distributed among parents of children from classes of Nursery and Kindergarten 1 to identify families in which both Mandarin and English were spoken. 71 families met the requirement and 33 of them consented to participation. However, 12 of them did not complete the tasks. The final sample included 21 typically developing children (13 girls) between 4;5 and 6;5 (*M*age = 5;6). Parental questionnaire showed that 10 children received regular exposure to Mandarin and English from birth, and the rest started exposure to Mandarin/English from birth and English/Mandarin between 3 and 36 months. All children heard Mandarin and English from one or more caregivers and/or older siblings in the home (nine of the children had older siblings), with different amount of exposure to the two languages (see the section "Measures"). Most of the children had never lived outside Singapore for over 3 months except for one child who had visited relatives in Malaysia frequently. According to the parents' observation, 42.9% ($n = 9$) of the children were balanced between the two languages, 38.1% ($n = 8$) were more proficient in English than in Mandarin, and the remaining 19% ($n = 4$) were more proficient in Mandarin than in English.

Mothers of the children were invited to a storytelling task performed in Mandarin and in English at the participants' own home. Maternal input was chosen to be examined because our language exposure questionnaire data (details below) showed that the mothers were the main caregiver of their child[4] and there are emerging research interests in the quality of input provided by bilingual mothers (e.g., Hoff

4 The proportions of maternal input in total language input were 4–24% in Mandarin and 1–16% in English.

TABLE 2 Descriptive statistics for the 21 child participants.

	Mean	SD	Range	IQR
Age (months)	66.14	6.73	53–77	10.5
Current amount of exposure (proportion)				
Mandarin	44.23%	0.13	26.54–78.45%	0.15
English	53.43%	0.13	21.55–73.46%	0.21
Cumulative length of exposure (years)				
Mandarin	2.15	0.89	0.64–4.06	1.28
English	2.21	0.99	0.69–4.13	1.87
Working memory[a]				
BRIEF-P (raw score)	24.2	5.25	17–35	8
BRIEF-P (t-score)	52.2	10.36	38–73	16
Mandarin proficiency				
MVST (raw score)	13.19	5.09	5–24	8
MVST (scaled score)	6.19	2.52	2–13	3
English proficiency				
PPVT (raw score)	85.14	23.51	47–125	44
PPVT (standard score)	97.71	14.69	73–126	25.5

IQR, interquartile range
[a]Calculated based on data of 20 children.

et al., 2020). 81% ($n = 17$) of the mothers held Bachelor's degrees or higher, suggesting mid to high socioeconomic status background. 57.1% ($n = 12$) and 28.6% ($n = 6$) considered themselves (near-)native in Mandarin and English, respectively. 38.1% ($n = 8$) and 42.9% ($n = 9$) rated themselves as fluent speakers of Mandarin and English, respectively. 95.2% ($n = 20$) of the mothers addressed their child in both Mandarin and English. Sixteen mothers (out of 21) completed the task in both languages. One mother who mostly spoke Mandarin to her child performed the task only in Mandarin.[5]

Measures

We collected information on the children's language exposure in addition to demographic information and language profiles of their main caregivers through a web-based interview with the parent(s). We measured the children's working memory, and language proficiency in Mandarin and English, using standardized assessment tools.[6] Participation was ascertained through parental consent forms. The study was approved by the Survey and Behavioural Research Ethics Committee of the Chinese University of Hong Kong. Descriptive statistics of background variables are given in Table 2.

Language exposure

We used a parental questionnaire (in the form of an excel file) modeled on the BiLEC (Unsworth, 2013) to estimate the children's relative amount of exposure to Mandarin and English concurrently and cumulatively. The parents (usually the mother) met members of the research team *via* the web conferencing software, Zoom Meetings. They answered questions about the child's CURRENT LANGUAGE EXPOSURE on a weekly basis including (i) hours of interaction and language spoken with each input provider in the home and friends and relatives on average weekday and at weekends, (ii) language and hours of school and after-school activities, and (iii) language and hours of the child's experience with media (e.g., TV, videos, books, and computer games). We calculated the proportion of time the child interacted with each input provider during waking hours and multiplied it with the percentage of Mandarin/English used by the respective input provider. The same applies to the calculation of the child's language exposure in media/school/after-school activities. We added up the figures to derive the child's relative amount of current exposure to Mandarin and English, respectively. For CUMULATIVE LENGTH OF EXPOSURE, parents recalled (i) the frequency at which each caregiver (and school-age older siblings, if any) in the home spoke Mandarin/English for each one-year period in the child's life, (ii) language use in daycare and/or school and/or out-of-school-care in these periods, and (iii) language use in the holidays. We averaged the frequency of Mandarin/English exposure at home for each period. We then calculated the proportion of time the child spent at home/daycare/school/out-of-school-care each year based on what is typical in Singapore and worked out the proportion of year with Mandarin/English exposure in each context. The estimates were summed up to obtain the cumulative exposure in each language. The range of current exposure in our sample is

5 It was the older sibling (10-year-old) who communicated with the child in English at home. The remaining 4 mothers were not available.

6 We also collected other individual difference measures which are not the focus of the current study and therefore not reported here.

26.5–78.4% in Mandarin ($M = 44.2\%$, $SD = 0.13$), and 21.5–73.5% in English ($M = 53.4\%$, $SD = 0.13$), with 16 of the bilingual children receiving a greater amount of input from English than Mandarin. Six of the children were also exposed to other languages, namely Cantonese, Hokkien, Teochew, and Japanese. Current exposure to other languages mostly accounted for less than 9% of the input except for one child (30.1%). The cumulative exposure in our sample was 0.64–4.06 years in Mandarin ($M = 2.15$, $SD = 0.89$), and 0.69–4.13 years in English ($M = 2.21$, $SD = 0.99$). Current and cumulative exposure were highly correlated in our sample for Mandarin (Pearson correlation, $r = 0.78$, $p < 0.001$) and English ($r = 0.76$, $p < 0.001$).

Working memory

We used the Behavior Rating Inventory of Executive Function-Preschool Version (BRIEF-P; Gioia et al., 2003) to measure WM in the preschoolers. BRIEF-P is a questionnaire completed by parents or teachers to reflect a child's executive functions in everyday environment, using a three-point problem-oriented symptom rating scale. It has been reported to correlate to a varying degree with performance-based executive function assessment results in preschool children (e.g., Espy et al., 2011; Garon et al., 2016; O'Meagher et al., 2018). It is thus a good alternative to directly assessing the children during the pandemic. For this study, we adopted the WM sub-score of BRIEF-P as a proxy measure for children's WM. In items relevant to WM, parents were asked to describe their child's capacity to hold information in mind for completing a task or making a response—for instance, forgetting directions, losing track of what they are doing in the middle of an activity, unable to finish describing an event, person or story, and forgetting what they are supposed to retrieve when instructed, etc. Parents of 20 (out of 21) children completed BRIEF-P during a virtual meeting with our research assistants, approximately 6 months before the administration of the elicitation tasks and other measures.

Language proficiency

We used two standardized tests, namely, Peabody Picture Vocabulary Test—Fourth Edition (PPVT-4; Dunn and Dunn, 2007) and the receptive vocabulary subtest of the Taiwan version of the Wechsler Preschool and Primary Scale of Intelligence—Fourth Edition (WPPSI-IV; Wechsler, 2013) to measure language proficiency in English and Mandarin, respectively.[7] In both tests, children were presented four colored pictures each time and their task was to select the one that matched the word they heard. PPVT-4 was administered by using digital tools from Q-global for teleassessment. WPPSI-IV Mandarin vocabulary subtest (MVST) was administered in accordance with guidelines from the test publisher for teleassessment (displaying the stimuli using a camera). A standard score between 85 and 115 on the English PPVT-4 scale and a scaled score between 7 and 12 on the MVST indicate that an examinee's raw score is within the average of the age-matched monolinguals in the respective normative sample. It is clear from Table 2 that our bilingual children were generally more advanced in English than in Mandarin.

Elicited narration task

Oral narratives were elicited remotely in real time with the picture sequences *Baby Birds* and *Baby Goats* from the Multilingual Assessment Instrument for Narratives (MAIN; Gagarina et al., 2012, 2015, 2019), which has been successfully used to elicit oral narratives in face-to-face settings from children speaking different languages including Mandarin (Sheng et al., 2020). The stories depict comparable character actions and emotions, and have parallel episodic structures. Both involve five characters that are familiar to young children: a mommy bird/a mommy goat, two baby birds/two baby goats, a cat/a fox, and a dog/a crow. Each story is made up of three episodes, with two pictures depicting an episode.

We adapted one of the PowerPoint templates of MAIN (Hamdani et al., 2021) for remote testing. The adaptations included the use of animation in place of videos to show the folding/unfolding of the picture sequences. The MAIN instructions were pre-recorded by two female fluent speakers of Mandarin and English respectively, following the MAIN manual (Gagarina et al., 2019; Luo et al., 2020). Each child was tested once in each language, with an interval of about 1 week between sessions. The order of language and stories was counterbalanced. Half ($n = 10$) of the children were tested in English first and Mandarin second and vice versa. Eleven children told *Baby Birds* in Mandarin and *Baby Goats* in English, and 10 told *Baby Goats* in Mandarin and *Baby Birds* in English.

All participants were individually tested in a quiet room at school. They were accompanied by a teacher, who provided technical assistance to the child. The teachers remained silent during the test so as not to disturb or distract the child. The child sat in front of a computer and met an experimenter based in Hong Kong *via* Zoom (illustrated in Figure 2). The experimenters (the third and fourth authors of this article) are fluent speakers of Mandarin and English but were posing as monolingual speakers of the languages, respectively, throughout the study to administer the Mandarin and English tasks separately. Test began following a short warm-up phase to establish rapport. The experimenter presented the PowerPoint using the share-screen-with-audio function of Zoom. The child viewed the shared screen in side-by-side mode, with the shared screen on the left and the video of the

7 Although PPVT-4 and WPPSI-IV were normed on native speakers of English in the U.S and native speakers of Mandarin Chinese in Taiwan, respectively, they are a pragmatic solution that can provide independent measures of our bilingual children's proficiency in English and Mandarin, given the lack of culturally appropriate/neutral standard tests that target bilingual children. Recall that the primary interest of the current study is not bilingual versus monolingual differences in absolute terms but relationships between variables within the bilingual group.

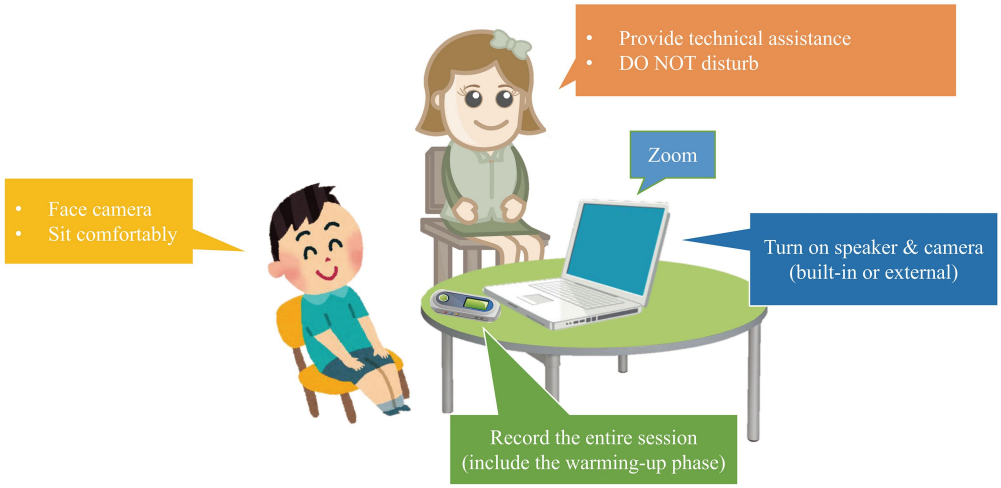

FIGURE 2
Illustration of remote elicitation of child narratives over an audio/video platform in real time.

experimenter on the right. Three envelopes in different colors appeared on the screen and the child was asked to choose one. Whichever was chosen, the same story was presented but this was unbeknownst to the child. The child was given some time to preview the entire picture sequences. Then the pictures were "folded" and reappeared on the screen, two at a time. The child was asked to tell the story to the experimenter. Previous studies found that the presence of shared access to the referent affected children's use of REs (e.g., Kail and Hickmann, 1992). To create the desired non-shared visual attention, the experimenters covered their eyes with their hands or a sheet of paper and made sure the child noticed it before the picture sequences were shown. The child was told to let the experimenter know when a given slide was done. By doing so, we hope to reduce the impact of screen sharing on children's referential strategies. The session was video-recorded by the experimenter using the built-in recording function in Zoom and audio-taped by the school teacher accompanying the child using a mobile phone at the same time. The on-site audio recording was to remedy for likely unstable internet connection and subsequent loss of signals during the Zoom calls. The transcription and coding (to be introduced below) were performed based on an edited version of the Zoom video recording, in which the soundtrack was replaced by the on-site audio recording. This apparatus and setup was first created in remote web-based data collection for the Child Heritage Chinese Corpus (Mai and Yip, in prep) in CHILDES (MacWhinney, 2000) and adopted in a series of similar studies by the team (e.g., Mai et al., in prep).

Recording home storytelling by mother

Participating mothers received a hardcopy of the picture sequences of the two MAIN stories (printed on A4 paper). They were asked to tell the stories to their child at home in the way they would normally do (illustrated in Figure 3). Both stories were told twice on different days, once in Mandarin and once in English. The order was determined freely by the mother. The two Hong Kong-based experimenters video-recorded the mother–child interaction with Zoom. They remained muted and invisible during the recording. Like storytelling by the children recorded in the school, additional on-site audio-recording was also obtained through the mother and edited into the video recording. It took around 5 min to complete the recording in each language.

Transcription and coding

Children's oral narratives and mother's home storytelling samples were transcribed verbatim in the CHAT-format (MacWhinney, 2000) and carefully checked. Transcription included non-verbal information relevant to referential choice such as pointing during mother–child interaction, which was captured by the video recordings.

Each reference to the story characters (excluding REs used in imagined dialogues between story characters[8]) was coded in terms of referential form, syntactic position (INTRO only), and discourse function, excluding unclear or unintelligible utterances.

Referential forms were first coded into different RE types based on Hickmann et al. (1996) and Jia and Paradis (2015): [Num-Cl-N] (Mandarin), [Indef./Def./Num. determiner-NP] (English), bare nouns (Mandarin), no determiner singular N (used as proper nouns in English, e.g., *Cat is so naughty*), demonstrative NPs,[9] kinship terms, complex NPs containing a possessor, a relative clause

8 The REs in imagined dialogue reflected the perspective of story characters rather than the perspective of the narrator or the listener.

9 When coding the Mandarin data, we excluded cases in which the demonstrative *nage* could be treated as discourse gap fillers (i.e., there was a long pause between the demonstrative and the nominal).

FIGURE 3
Illustration of remote recording of mother–child interactions over an audio/video platform in real time.

or adjectival modification, personal pronouns, demonstratives, null forms, and non-specific lexical items (e.g., *someone*).

Syntactic position of REs in INTRO contexts was coded as preverbal or postverbal. Cases in which position is irrelevant or cannot be determined (e.g., labeling without predication) were coded as unanalyzable and excluded from the analysis, as in Hickmann and Liang (1990).

Discourse function was coded largely following Serratrice (2007), with reference to Colozzo and Whitely (2014). The unit of analysis is "clause" defined by the presence of a verbal predicate (Serratrice, 2007). The verbal predicates are mainly verbs and may include adjectives in Mandarin. INTRO is the first mention of a character. Re-INTRO involves topic shift across adjacent clauses. To be coded as Re-INTRO, an RE must meet one of the following criteria: (i) a subject/object argument referring to a previously identified referent which has not been mentioned in the immediately preceding clause; (ii) a subject argument that has been mentioned in the adjacent clause as a non-subject (e.g., an object or an adjunct); or (iii) the reference shifts from two or more characters together to only one of these characters (and vice versa). The participating mothers often interacted with their child by asking questions and discussing the plots when performing the home storytelling task. Child utterances which were relevant to the thematic progress of the story were treated as part of the discourse and taken into consideration when coding the discourse functions of REs in the mother data. Examples of the coding are given in the Supplementary Material.

To assess intercoder reliability, the data were coded independently by the first author (C1) and the third and fourth authors (C3 and C2), all of whom were Mandarin-English bilinguals: C1 coded all the data, C2 coded the English child data, and C3 coded the Mandarin data and the English mother data. The agreement rate (i.e., the percentage of items with consistent coding between coders out of the total number of coded items) was 99.48% between C1 and C2 and 99.81% between C1 and C3. All inconsistencies were discussed among the coders until consensus was reached.

Results

In total, the narratives yielded 248 REs (INTRO 82, Re-INTRO 166) in child Mandarin, 257 REs (INTRO 80, Re-INTRO 177) in child English, 777 REs (INTRO 114, Re-INTRO 663) in mother Mandarin, and 658 REs (INTRO 111, Re-INTRO 547) in mother English. For expository convenience, we further categorized the nominals into indefinite nominals and definite/identifiable nominals based on their expected interpretation in the target grammar (see Table 1). The child data were subject to Chi-square tests to rule out potential effects of story and testing order by comparing occurrences of indefinite nominals, definite/identifiable nominals, pronominals, and null forms. Results showed that the two stories elicited comparable number of REs for INTRO [$\chi^2(3, N = 162) = 1.614$, $p = 0.656$] and Re-INTRO [$\chi^2(3, N = 343) = 2.945$, $p = 0.400$], and there was no

TABLE 3 Descriptive statistics of the referential expressions (REs) produced by the Mandarin-English bilingual children (n =21) and their mothers (n =17) to introduce and reintroduce characters (INTRO, Re-INTRO) in Mandarin.

	INTRO		Re-INTRO	
	Child	Mother	Child	Mother
[Num-Cl-N]	29.27% (24)	54.39% (62)	10.84% (18)	8.6% (57)
Bare noun	18.29% (15)	9.65% (11)	15.66% (26)	38.91% (258)
Complex NP	10.98% (9)	24.56% (28)	8.43% (14)	21.87% (145)
Kinship term	2.44% (2)	0	1.2% (2)	6.64% (44)
Demonstrative NP	35.37% (29)	4.39% (5)	48.19% (80)	10.26% (68)
Demonstrative	0	7.02% (8)	0	1.21% (8)
Personal pronoun	2.44% (2)	0	12.65% (21)	8.14% (54)
Null form	1.22% (1)	0	3.01% (5)	4.37% (29)

TABLE 4 Pre/postverbal positioning of referential expressions (REs) in INTROs in Mandarin-English bilingual children and their mothers: Mandarin.

	Child (n = 21)		Mother (n = 17)	
	Preverbal	Postverbal	Preverbal	Postverbal
Overall	71.95% (59)	28.05% (23)	44.44% (48)	55.56% (60)
[Num-Cl-N]	45.83% (11)	54.17% (13)	8.62% (5)	91.38% (53)
Bare nouns	86.67% (13)	13.33% (2)	90.9% (10)	9.09% (1)
Complex NP	66.67% (6)	33.33% (3)	76.92% (20)	23.08% (6)
Kinship term	50% (1)	50% (1)	0	0
Demonstrative NP	93.1% (27)	6.9% (2)	100% (5)	0
Demonstrative	0	0	100% (8)	0
Personal pronoun	50% (1)	50% (1)	0	0
Null form	0	100% (1)	0	0

difference between English-first and Mandarin-first groups [INTRO $\chi^2(3, N = 162) = 5.647$, $p = 0.130$; Re-INTRO $\chi^2(3, N = 343) = 0.526$, $p = 0.914$].

We analyzed the distribution of REs in each participant for each language by calculating their percentages among the total number of REs for INTRO and Re-INTRO, respectively. This is to assess the similarities and differences in the use of REs between Mandarin and English, and between bilingual children's production and maternal input. For each discourse function, we examined whether bilingual children's production of different types of REs correlated with maternal input (in terms of structure frequency). We also implemented generalized linear mixed-effects logistic regression models to investigate the effects of linguistic, input, and cognitive factors and their interactions in bilingual reference production. Most of the statistical tests were run using IBM SPSS Statistics Version 26 except for the mixed-effects analyses, for which we used the R package *lme4* (Bates et al., 2015) in the statistical program R (version 3.5.2, R Core Team, 2018). The Shapiro–Wilk test showed that some of the variables were not

TABLE 5 Descriptive statistics of the referential expressions (REs) produced by the Mandarin-English bilingual children (n =21) and their mothers (n =16) to introduce and reintroduce characters (INTRO, Re-INTRO) in English.

	INTRO		Re-INTRO	
	Child	Mother	Child	Mother
[indef. determiner-NP]	21.25% (17)	36.04% (40)	0	2.19% (12)
[num. determiner-NP]	11.25% (9)	16.22% (18)	4.52% (8)	3.11% (17)
[def. determiner-NP]	62.5% (50)	26.13% (29)	85.31% (151)	70.38% (385)
No determiner singular N	1.25% (1)	0.9% (1)	0.56% (1)	0.37% (2)
Complex NP	2.5% (2)	9.91% (11)	2.26% (4)	10.05% (55)
Kinship term	0	0	0	1.65% (9)
Demonstrative NP	1.25% (1)	5.41% (6)	0.56% (1)	2.38% (13)
Demonstrative	0	2.7% (3)	0	0.91% (5)
Personal pronoun	0	0.9% (1)	4.52% (8)	7.5% (41)
Null form	0	0	2.26% (4)	1.46% (8)
Non-specific lexical item	0	1.8% (2)	0	0

normally distributed. Therefore, results of nonparametric tests will be reported unless indicated otherwise. To calculate the *post hoc* power analysis for mixed models, we employed the *simr* package (Green and MacLeod, 2016) in R. Tables 3–6 present the distribution of REs used to introduce and reintroduce characters in Mandarin and English, respectively.

Referential choice for introduction of characters

Mandarin

Compared to the mothers, the bilingual children produced more demonstrative NPs (35.37% vs. 4.39%; Mann–Whitney test, $U = 84$, $z = -3.01$, $p = 0.003$), fewer [Num-Cl-N] (29.27% vs. 54.39%; $U = 87$, $z = -2.706$, $p = 0.007$), and a lower rate of complex NPs (10.98% vs. 24.56%; $U = 97$, $z = -2.051$, $p = 0.012$) for INTRO in Mandarin. They produced more bare nouns (18.29% vs. 9.65%) than the mothers, though the difference was non-significant ($p = 0.484$). Pronominals and null forms were used infrequently (0–7.02%) for INTRO by the children and the mothers. We performed bivariate correlation tests to find out the relations between children's production and maternal input in terms of structural frequency. A significant correlation was found for the use of bare nouns in Mandarin INTRO contexts (Spearman's rank correlation, $r_s = 0.482$, $p = 0.05$).

The INTROs in child Mandarin were more often preverbal than postverbal (71.95% vs. 28.05%) (Wilcoxon signed-rank tests, $z = -2.541$, $p = 0.011$). Pre- and postverbal INTROs (44.44% vs. 55.56%) were almost equally distributed in the maternal input ($p = 0.477$). The children produced significantly fewer postverbal INTROs than the mothers ($U = 95$, $z = -2.45$, $p = 0.014$).

About half of the [Num-Cl-N] were preverbal in the bilingual children (54.17%), whereas most [Num-Cl-N] appeared postverbally in the maternal input (91.38%). Bare nouns were

TABLE 6 Pre/postverbal positioning of referential expressions (REs) in INTROs in Mandarin-English bilingual children and their mothers: English.

	Child ($n = 21$)		Mother ($n = 16$)	
	Preverbal	**Postverbal**	**Preverbal**	**Postverbal**
Overall	68.42% (52)	31.58% (24)	48.91% (45)	51.09% (47)
[indef. determiner-NP]	41.18% (7)	58.82% (10)	37.84% (14)	62.16% (23)
[num. determiner-NP]	44.44% (4)	55.56% (5)	21.43% (3)	78.57% (11)
[def. determiner-NP]	85.11% (40)	14.89% (7)	68.18% (15)	31.82% (7)
No determiner singular N	0	0	100% (1)	0
Complex NP	0	100% (2)	71.43% (5)	28.57% (2)
Demonstrative NP	100% (1)	0	80% (4)	20% (1)
Demonstrative	0	0	33.33% (1)	66.67% (2)
Personal pronoun	0	0	100% (1)	0
Null form	0	0	0	0
Nonspecific lexical item	0	0	50% (1)	50% (1)

mostly preverbal (child 86.67%, mother 90.9%); so were demonstrative NPs (child 93.1%, mother 100%).

English

For INTRO in English, our bilingual children differed from the mothers in producing more [Def. determiner-NP] (62.5% vs. 26.13%; $U = 62$, $z = -3.272$, $p = 0.001$), fewer [Indef. Determiner-NP] (21.25% vs. 36.04%; $U = 100$, $z = -2.136$, $p = 0.033$) and a lower rate of complex NPs (2.5% vs. 9.91%; $U = 104$, $z = -2.510$, $p = 0.012$). While [Num. determiner-NP] was used occasionally (child 11.25%, mother 16.22%), the use of the other RE types was rare (child 0–1.25%, mother 0–5.41%). There was no significant mother–child correlation regarding structural frequency of REs in English INTRO contexts ($ps > 0.4$).

The INTROs in English were mostly preverbal as opposed to postverbal (68.42% vs. 31.58%) in the bilingual children ($z = -2.583$, $p = 0.01$) and almost equally distributed between the pre- and postverbal positions in the mothers (48.91% vs. 51.09%, $p = 1$). The children produced significantly fewer postverbal INTROs than the mothers ($U = 96.5$, $z = -2.212$, $p = 0.027$).

[Indef. determiner-NP] and [Num. determiner-NP] were often postverbal in the bilingual children (Indef. 58.82%, Num. 55.56%), and mostly postverbal in the mothers (Indef. 62.16%, Num. 78.57%). By contrast, most [Def. determiner-NP] appeared preverbally (child 85.11%, mother 68.18%).

Referential choice for re-introduction of characters

Mandarin

For Re-INTRO in Mandarin, demonstrative NPs were used most frequently by the bilingual children (48.19%), followed by bare nouns (15.66%), personal pronouns (12.65%), and complex NPs (8.43%). This contrasts with the maternal input, in which bare nouns (38.91%) and complex NPs (21.87%) were used more frequently than demonstrative NPs (10.26%) and personal pronouns (8.14%). The child–mother differences were significant with demonstrative NPs ($U = 83.5$, $z = -2.804$, $p = 0.005$), bare nouns ($U = 60$, $z = -3.551$, $p < 0.001$), and complex NPs ($U = 47$, $z = -3.924$, $p < 0.001$). [Num-Cl-N] was used occasionally by the children (10.84%) and the mothers (8.6%). The use of demonstratives and null forms for Re-INTRO was infrequent in Mandarin (0–4.37%). A positive mother–child correlation was found with the frequency of demonstrative NPs in Mandarin Re-INTRO contexts ($r_s = 0.548$, $p = 0.023$). The bilingual children's use of demonstrative NPs increased as the frequency of demonstrative NPs in the maternal input increased.

English

Our children's Re-INTROs in English were patterned after the maternal input: [Def. determiner-NP] occurred the most frequently (child 85.31%, mother 70.38%), while the other REs were infrequent (child 0–4.52%, mother 0–7.5%), except that the mothers showed occasional use of complex NPs (10.05%). There was no significant mother–child correlation regarding structural frequency of specific types of REs in English Re-INTRO contexts ($ps > 0.2$).

Multifactorial modeling

We generated four mixed-effects logistic regression models. Two modeled the bilingual children's reference production in Mandarin (Model 1) and English (Model 2), and the others modeled their reference production for INTRO (Model 3) and Re-INTRO (Model 4). In these models, the referential choice was entered as binary data and participants were treated as a random effect.[10] Categorical factors were sum-coded (i.e., −0.5 and 0.5)

10 Only random intercepts were included. Models with random slopes either failed to converge or were not a better fit of the data as indicated by anova() comparisons. The patterns of results did not change in fuller models.

TABLE 7 Results from a mixed-effects logistic regression model on Mandarin-English bilingual children's (n =20) choice of definite nominals vs. other REs in the Mandarin oral narrative task (229 observations).

Predictor	Estimate	SE	z value	p value
Intercept	0.97	0.199	4.868	< 0.001***
Discourse function (Re-INTRO vs. INTRO)	0.134	0.163	0.825	0.409
Cumulative length of exposure (Mandarin)	−0.101	0.238	−0.425	0.671
Working memory	0.024	0.043	0.558	0.577
Mandarin proficiency (MVST raw scores)	0.026	0.044	0.585	0.559

***$p < 0.001$.

TABLE 8 Results from a mixed-effects logistic regression model on Mandarin-English bilingual children's (n =20) choice of definite nominals vs. other REs in the English oral narrative task (243 observations).

Predictor	Estimate	SE	z value	p value
Intercept	1.956	0.298	6.568	< 0.001***
Discourse function (Re-INTRO vs. INTRO)	0.412	0.251	1.639	0.101
Cumulative length of exposure (English)	−0.261	0.326	−0.8	0.424
Working memory	0.011	0.053	0.213	0.831
English proficiency (PPVT raw scores)	−0.041	0.016	−2.639	0.008**
Discourse function × English proficiency	0.028	0.012	2.301	0.021*

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

and continuous variables were mean centered (by subtracting the mean from the value). Since a higher score obtained in the BRIEF-P assessment suggests weaker executive functions, the working memory scores were reversed by multiplying "-1" after the mean-centering procedure to align with other variables. Two-way interactions between predictor variables were included if they significantly improve model fit as measured by Akaike Information Criterion.

For Models 1 and 2, the dependent variable was the choice between definite/identifiable nominals and others. As fixed effects, we entered (i) Discourse function as a two-level factor (INTRO, Re-INTRO), (ii) continuous predictors including Cumulative length of exposure[11] in Mandarin/English, Working memory (raw scores), and language proficiency (MVST/PPVT raw scores). The interaction between discourse function and English language proficiency was included in Model 2 as it significantly improved model fit. Tables 7, 8 show the results.

Mandarin (Model 1)

No significant effect was found ($ps > 0.4$).

English (Model 2)

There was a significant main effect of language proficiency ($\beta = -0.041$, $SE = 0.016$, $z = -2.639$, $p = 0.008$; *post hoc* power = 84.4%), which was qualified by discourse function ($\beta = 0.028$, $SE = 0.012$, $z = 2.301$, $p = 0.021$; *post hoc* power = 94.7%). That is, the production of definite/identifiable nominals (infelicitous) in INTRO contexts decreased as proficiency increased, while the probability of definite/identifiable nominals (felicitous) in Re-INTRO contexts was similarly high across proficiency. No effect of cumulative exposure to English was found ($p = 0.424$).

For Models 3 and 4, the dependent variable was indefinite nominals vs. others and definite/identifiable nominals vs. others, respectively. As fixed effects, we entered (i) Language as a two-level factor (English, Mandarin), (ii) continuous predictors including Structural frequency (of indefinite nominals for INTRO or definite/identifiable nominals for Re-INTRO) in the maternal input, Working memory (raw scores), Relative cumulative exposure (subtracting the child's cumulative length of exposure to Mandarin from her/his cumulative length of exposure to English), and Age. The interaction between structural frequency and working memory was included in both models as it significantly improved model fit. Tables 9, 10 show the results.

INTRO (Model 3)

There was a marginally significant age effect ($\beta = 0.083$, $SE = 0.046$, $z = 1.814$, $p = 0.07$; *post hoc* power = 49%), suggesting a trend in more indefinite nominals (felicitous for INTRO) with increasing age. Working memory interacted with structural frequency ($\beta = 0.674$, $SE = 0.220$, $z = 3.064$, $p = 0.002$; *post hoc* power = 98.2%): children with stronger working memory capacity and higher frequency of indefinite nominals in the maternal input were more likely to produce indefinite nominals in introducing characters. Whether the children received more exposure to English than to Mandarin did not show any significant effects on the production of indefinite nominals in INTRO contexts ($p = 0.287$).

11 The same pattern of results was obtained when current exposure was entered. This holds for the other mixed analyses. We therefore modelled the data with cumulative exposure throughout the paper for consistency.

TABLE 9 Results from a mixed-effects logistic regression model on Mandarin-English bilingual children's (n=17) choice of indefinite nominals vs. other REs in INTRO contexts (128 observations).

Predictor	Estimate	SE	*z* value	*p* value
Intercept	−1.027	0.268	−3.836	<0.001***
Language (Mandarin vs. English)	−0.209	0.219	−0.958	0.338
Structural Frequency	−1.519	1.1	−1.382	0.167
Working memory	0.06	0.063	0.953	0.341
Relative cumulative length of exposure	1.29	1.213	1.064	0.287
Age	0.083	0.046	1.814	0.07
Structural frequency × Working memory	0.674	0.22	3.064	0.002**

***$p < 0.001$; **$p < 0.01$.
Structural frequency, frequency of indefinite nominals in the maternal input; Relative cumulative length of exposure, English-Mandarin differences in cumulative length of exposure.

TABLE 10 Results from a mixed-effects logistic regression model on Mandarin-English bilingual children's (n =17) choice of definite/identifiable nominals vs. other REs in Re-INTRO contexts (289 observations).

Predictor	Estimate	SE	z value	p value
Intercept	1.819	0.268	6.777	<0.001***
Language (Mandarin vs. English)	−0.694	0.198	−3.508	<0.001***
Structural frequency	5.323	2.921	1.822	0.068
Working memory	0.031	0.053	0.584	0.559
Relative cumulative length of exposure	−0.567	1.082	−0.524	0.6
Age	0.019	0.04	0.485	0.628
Structural frequency × Working memory	1.224	0.616	1.986	0.047*

***$p < 0.001$; *$p < 0.05$.
Structural frequency, frequency of definite/identifiable nominals in the maternal input; Relative cumulative length of exposure, English-Mandarin differences in cumulative length of exposure.

Re-INTRO (Model 4)

Language was a significant predictor ($\beta = -0.694$, $SE = 0.198$, $z = -3.508$, $p < 0.001$; *post hoc* power = 96.7%). The bilingual children produced more definite/identifiable nominals (felicitous for Re-INTRO) in English than in Mandarin. Structural frequency showed a marginally significant main effect ($\beta = 5.323$, $SE = 2.921$, $z = 1.822$, $p = 0.0068$; *post hoc* power = 59.5%), and interacted with working memory ($\beta = 1.224$, $SE = 0.616$, $z = 1.986$, $p = 0.047$; *post hoc* power = 56.7%). In other words, children with stronger working memory capacity produced more definite/identifiable nominals for Re-INTRO with increasing frequency of these nominals in the maternal input. The production of definite/identifiable nominals in Re-INTRO contexts did not change as a function of relative cumulative exposure ($p = 0.6$).

Discussion

Summary of main findings

The current study investigated the relationship between reference production on the one hand, and linguistic, input, and working memory on the other by examining referential choice in 4–6-year-old Singaporean Mandarin-English bilingual children through a bilingual elicited narration task, supplemented by a battery of language proficiency, input and cognitive measures.

Our first research question concerns bilingual children's referential choice for INTRO and Re-INTRO contexts. The results showed that our bilingual children overused definite nominals for INTRO in both Mandarin and English. The use of indefinite nominals in INTRO contexts improved as a function of language proficiency with English but not with Mandarin. Although we did not include monolingual groups in this study, below we make comparisons drawing on the trends and patterns in the English/Mandarin monolingual preschoolers reported in Hickmann et al. (1996) (hereafter HHRL) and 5-year-old Mandarin monolinguals in Wu et al. (2015) (hereafter WHZ) in terms of INTRO contexts. Both HHRL and WHZ tested children's reference production using elicited narration tasks similar to our study.

In English, our bilingual children produced more [Def. determiner-NP] (62.5%) than [Indef. determiner-NP] (21.25%), similar to the English monolingual peers (HHRL 62% vs. 25%). Different patterns of results were observed in Mandarin, however. Our bilingual children produced more demonstrative NPs (35.37%) than [Num-Cl-N] (29.27%) and bare nouns (18.29%), while Mandarin monolingual preschoolers used [Num-Cl-N] (HHRL 50%, WHZ 47–73%) and bare nouns (HHRL 34%, WHZ 27–49%) more frequently than demonstrative NPs (HHRL 17%, WHZ 0–4%). First mentions were more often preverbal than postverbal in both our bilingual children (68.42% vs. 31.58%) and the English monolingual children (HHRL around 70% vs. 30%). The same holds in Mandarin, though the difference in proportions

seems to be larger in our bilingual children (72% vs. 28%) than in the monolingual children (HHRL 56% vs. 44%, WHZ 64.68% vs. 35.32%). This is partially consistent with Jia and Paradis (2015), who reported no preference for the postverbal position in heritage Mandarin speakers, despite that first mentions are typically postverbal in Mandarin.

Re-INTRO constitutes felicitous contexts for definite/identifiable nominals. As expected, definite/identifiable nominals were more frequent in Re-INTRO contexts than INTRO contexts, especially in English, as revealed by the mixed-effects analysis. The higher rate of definite/identifiable nominals in English than in Mandarin is expected and probably attributable to Mandarin-English differences independent of bilingualism, as the same pattern was found in the monolingual children (Mandarin 69.2%, English 84.4%) in Chen and Lei (2012). Our bilingual children produced a higher frequency of demonstrative NPs (80 out of 140 nominals, 57.14%) than the 6–9-year-old typically developing Mandarin monolingual children in Sah (2018) (24 out of 276 nominals, 8.7%) for Re-INTRO. This echoed the bilingual-monolingual differences in demonstrative use between heritage Mandarin speakers and homeland speakers reported in Aalberse et al. (2017) and Mai et al. (2021).

The results overall show that our bilingual children were sensitive to differential uses of REs in INTRO and Re-INTRO contexts while overproducing definite nominals for INTRO. Meanwhile, they showed specific uses of REs in Mandarin (partially) consistent with previous findings, including an increase in the use of demonstrative NPs and the prevalence of preverbal INTROs, which will be returned to in the next section.

Our second research question examines differences and correlation between children's reference production and maternal input in terms of structural frequency, and the role of input in bilingual reference production. We performed a qualitative analysis of the REs in the children and their mothers. The referential choice in our bilingual children generally patterned with that by their mothers except for two child–mother differences:[12] the children (i) produced indefinite nominals less frequently and preferred the preverbal position in INTRO contexts, and (ii) employed overt marking to code definiteness more frequently in Mandarin. Correlation analyses revealed positive relations between the children and the mothers in terms of structural frequency of (i) bare nouns for INTRO in Mandarin, and (ii) demonstrative NPs for Re-INTRO in Mandarin. Mixed-effects analyses showed that the frequency of felicitous REs produced by the children increased with a higher frequency of felicitous REs in the maternal input, modulated by working memory in both INTRO and Re-INTRO contexts. These results are consistent with the observation in Paradis and Navarro (2003), suggesting that our bilingual children were sensitive to the structural frequency in the input, which impacted on the patterns of their reference use. Amount of language exposure turned out to show no predicting effect on reference production, which is inconsistent with previous studies (e.g., Jia and Paradis, 2015). We will return to this in the section "Role of input in bilingual reference production."

Our third research question investigates the effect of working memory on bilingual reference production. As mentioned, there was a modulating effect of working memory on the mother–child association in the production of felicitous REs regardless of discourse context and language. Children with stronger working memory capacity and more frequent felicitous REs in the maternal input were better able to produce felicitous REs. These results are in line with previous evidence of working memory influencing child reference production (e.g., Torregrossa, 2017; Serratrice and De Cat, 2020).

Specific uses of REs in bilingual reference production

Compared to maternal input, our bilingual children under-produced indefinite nominals in INTRO contexts in English and Mandarin as expected. As proficiency increased, they produced a higher frequency of indefinite nominals for INTRO in English but not in Mandarin. This suggests that linguistic properties involving information structure and discourse such as REs in Mandarin could be particularly vulnerable in bilingual grammars, consistent with existing patterns in other bilingual populations (e.g., Mai and Deng, 2019).

Our children showed non-adult-like preference for the preverbal position when mentioning new referents. This is shown in (7), in which a more appropriate structure in Mandarin is SV inversion (i.e., *ranhou lai le yi-ge huli* "then came a fox").

7. Ranhou yi-ge huli lai le.

> then one-CL fox come LE
> Intended: "Then a fox came." (JL, 5;11)

One explanation for the "preverbal" preference is young children's preference of novelty. While the clause-initial position is typically associated with highly accessible referents (e.g., already mentioned, hence activated, and accessible) in adult

12 Our mothers produced fewer [Num-Cl-N] (54.39%) for INTRO in Mandarin than the adult controls in HHRL (81%) and WHZ (79–94%). Likewise, they produced fewer [Indef. determiner-NP] (36.04% vs. 76%) and more [Def. determiner-NP] (26.13% vs. 9%) for INTRO in English than the adults in HHRL. This is because our mothers told the stories during shared reading activities with their child (we deliberately chose shared reading to capture natural mother–child interaction) and the shared access to the stories might have influenced the mothers' referential choice. The adults in HHRL and WHZ, however, performed the task in the absence of shared knowledge. Recall that some of our mothers were non-native speakers of Mandarin/English. It is possible that the kind of Mandarin/English they spoke differed from that of native speakers. We leave this for future research.

speech, it may also be associated with novelty and change, resulting in new information being mentioned first (Bock et al., 2004). It has been shown that young children organize their sentences prioritizing novelty rather than accessibility, preferring to highlight new information first (Chen and Narasimhan, 2018; Chen et al., 2020). The preverbal preference observed in our study is consistent with these studies. Another possibility may lie in the differences in focus-marking between the two languages in general: Mandarin relies heavily on word order and syntactic constructions for focus-marking, whereas in English focus has a systematic manifestation *via* pitch accent, with less reliance on word order variation for realization (Chen et al., 2016). We conjecture that sustained exposure to English might have weakened the association between newness and postverbal positions in bilingual Mandarin grammars. The extent to which the preverbal preference is influenced by focus-marking of English needs further investigation.

Our expectation that bilingual children would show a high frequency of demonstrative NPs is confirmed. As shown in (8–9), our bilingual children frequently used the demonstrative *na* "that" plus a classifier (e.g., *na-ge*) to overtly mark definiteness, which is semantically redundant in Mandarin but well explained if *na* "that" was reanalyzed as the definite article *the* in English.

8. Ranhou na-ge gou yao zhua na-ge mao.

> then that-CL dog want catch that-CL cat
> "Then the dog wanted to catch the cat." (GX, 5;7)

9. Na-ge hei niao zhui zhe na-ge huli.

> that-CL black bird chase ASP that-CL fox
> "The black bird was chasing the fox." (LJ, 5;11)

Demonstratives in Mandarin are akin to the definite article in English in situations such as noncontrastive anaphoric reference and restrictive relative clauses (Chen, 2004). The obligatory use of the definite article in English might have trigged the search for an equivalent morpheme in Mandarin. Another possibility may be related to tolerance of "redundancy" due to a general effect of bilingualism—for example, the need to deal with higher processing load (Sorace and Filiaci, 2006; Sorace et al., 2009). It has been observed that bilingual children tend to be "over-explicit" in reference production, regardless of language combinations. They overused overt subjects/objects in contexts where a null form or clitic would be more appropriate (Paradis and Navarro, 2003; Belletti et al., 2007), and produced full nouns under circumstances where the use of pronominals is expected (Torregrossa et al., 2021). It follows that cross-linguistic influence and a general effect of bilingualism may jointly underlie the increase in the production of demonstrative NPs as observed in our data. Further research is needed to distinguish between the causes by comparing the use of demonstrative NPs in bilingual children acquiring two article-less languages. If those children also show over-reliance on overt marking of definiteness as our children did, it is likely due to a general effect of bilingualism.

Role of input in bilingual reference production

How input shapes language development is a question that features prominently in language-acquisition research. Recent studies have addressed this question by using various measures of linguistic input to predict children's language proficiency (e.g., Place and Hoff, 2016) or by investigating differential relations between input and acquisition of different linguistic structures (e.g., Paradis et al., 2014; Unsworth, 2014). The present study adds to the existing literature by presenting additional evidence regarding effects of different aspects of input in dual-language environments on bilingual reference production.

Our results highlight the robust correlation between structural frequency of REs in the input and patterns of reference use in 4–6-year-old Mandarin-English bilingual children. Maternal input positively correlated with children's production of specific types of REs, and importantly, structural frequency in the input interacted with working memory in predicting the bilingual children's production of felicitous REs (indefinite nominals for INTRO and definite/identifiable nominals for Re-INTRO) across languages. Thus, the results provide further support for Paradis and Navarro's (2003) observation that structural frequency in the input may be another source contributing to variation in children's referential choice. Note that REs produced by the children and the parents were collected and measured separately in different recording sessions. Although the mother told the stories in the presence of the child at home, the child told the stories in the kindergarten without the mother. This effectively reduces the possibility that mother–child associations in RE production are merely temporary adaptation effects between conversation interlocutors in general. Rather, the associations truly reflect the role of input on the acquisition outcomes in a longer term.

Recall that the production of indefinite nominals in Mandarin INTRO contexts is particularly challenging for our bilingual children. The input that the children received played a role here. As suggested by the structural frequency effect that was modulated by working memory, children who heard a higher frequency of indefinite nominals in the input and had stronger working memory capacity were better able to produce them. Following this, insufficient cues instantiating felicitous REs in the input may hamper children's development of referential abilities. Take bare nouns for example. It turned out that the mothers seldom produced bare nouns in INTRO contexts (9.65%) and when they did, most of the bare nouns they produced were preverbal (90.9%). The predominance of preverbal bare nouns over postverbal ones in INTRO contexts is unexpected since the opposite is believed to be the norm in Mandarin. Interestingly, in Wu et al. (2015), the adult controls who were university students speaking the homeland variety of Mandarin as the native language

also produced up to 41% of bare nouns in the preverbal position in INTRO contexts. The empirical evidence from the current study and Wu et al. (2015) both point toward a less significant tendency for bare nouns to appear postverbal in reference to new entities in Mandarin, compared to what was described in the theoretical literature (e.g., Cheng and Sybesma, 1999). Whether this discrepancy can be explained by contact-induced variation and change in Mandarin [e.g., contact influence from English which does not employ word order to mark (in)definiteness] awaits further investigation. In either case, ambiguity naturally arises in the input regarding the interpretation of bare nouns in pre- and postverbal positions from the perspective of acquisition. The input could be even less robust in our bilingual children than that of monolinguals, as the relevant amount of data in the input is reduced relative to monolingual children. Under such circumstances, it would be difficult for bilingual children to associate postverbal bare nouns with indefiniteness and preverbal ones with definiteness. This was borne out in our study, with most of the bare nouns for INTRO (86.67%) being preverbal in our children. Thus, our study suggests that less robust input with insufficient frequency of relevant structures would render REs more vulnerable for acquisition.

For now, we found no significant main effects of amount of language exposure on bilingual reference production. This appears to contradict the finding of Jia and Paradis (2015) who reported a significant effect of age of arrival on first mention abilities in 6–10-year-old heritage Mandarin children in Canada. It should be noted, however, that their children arrived at Canada at a rather young age (24 months on average) and developed bilingualism in one context-one language environment. Importantly, the effect of age of arrival was only observed in children who attended English-only schools (HL-ENG group), rather than in those who attended Mandarin-English bilingual public schools (HL-BIL group). In the current study, our bilingual children had been living in Singapore since birth, receiving exposure to both languages in diverse contexts. They were attending a Mandarin-English bilingual program in kindergarten with relatively balanced distribution of exposure between the two languages. Given this, the null effect of amount of language exposure in the current study (and perhaps the HL-BIL group in Jia and Paradis (2015)) may well result from the threshold effect of language exposure and the potentially non-linear nature between language exposure and language outcome (Pearson, 2007; Thordardottir, 2014). That is, the amount of input that our bilingual children received might have passed a certain amount above which increases in exposure would not add to performance in reference production.

The above said, it could be that the true effect of amount of language exposure in bilingual reference production will only emerge in a more focused design with stronger statistical power and more sensitive experimental tools. Following up on our behavioral findings, future research may further assess the effect of amount of exposure to evaluate this possibility by investigating a larger sample of children with a wide array of language dominance profiles.

Role of cognitive skills in bilingual reference production

In our study, working memory did not appear to make a significant individual contribution. Nevertheless, our results suggest that strong working memory capacity is particularly beneficial for RE acquisition among children who received input containing a higher frequency of felicitous nominals. This is consistent with studies that showed individuals with better working memory abilities are more efficient in attending to and decoding various features in the input (e.g., Sunderman and Kroll, 2009; Indrarathne and Kormos, 2018). Better working memory may assist in keeping information active for further processing and retaining it in the long-term memory, which expedites the retrieval of representations and extend the scope of attention (Martini et al., 2015), but this happens on the condition that there are sufficiently frequent cues in the input for the child to process.

The modulating effects of working memory are also in line with suggestions that better working memory helps bilingual children store, monitor, and update the addressee's perspective in their mind (De Cat, 2015; Serratrice and De Cat, 2020). As mentioned in the section "Working memory and reference production," reference to characters in storytelling tax working memory. Referential choice for INTRO is guided by the speaker's presupposition about the listener's knowledge, and Re-INTRO requires monitoring not only the knowledge state but also the attentional state of the listener (whether the character of concern is the attentional focus of the listener) to keep track of characters who are moving in and out of the attentional foreground and update the discourse model accordingly. In this sense, our result also aligns with previous research in which cognitive effects are shown to be pronounced in more complex working memory tasks (e.g., Morales et al., 2013; Blom et al., 2014).

Practical implications of the modulating effects of working memory on the mother–child association in reference production are two-fold: In multi-factorial predictive models of bilingual acquisition, pinpointing the role of working memory and its interaction with linguistic and input factors in RE acquisition facilitates more accurate predictions and expectations on the language developmental outcomes in bilingual children, given that RE is a prominent and challenging aspect in language. On the other hand, in intervention programs for bilingual children, pedagogical and educational effort can be made to utilize the positive and potentially reciprocal relation between working memory and language learning in order to promote mutual benefits for both sides. Several studies have shown that after training and intervention, working memory can improve and enhance language learning in children (see review in Archibald, 2017). In the opposite direction, there is also evidence of significant improvements in working memory after intervention targeting language skills such as phonological awareness skills (van

Kleeck et al., 2006), vocabulary, and morphosyntax (Ebert, 2014) in preschool and school-age children with specific language impairment.

Conclusion

This study investigated 4–6-year-old Mandarin-English bilingual children's reference production, and its relationship with linguistic, input, and cognitive factors. It is the first study of narrative production that has included transcript-based analysis of the maternal input available to preschool bilingual children captured through mother–child interactions. The current study is also one of the few studies that have elicited child and adult discourse remotely online using videoconference-based methods supplemented by on-site audio-recording.

Our data showed prolonged development of indefinite nominals to introduce a new referent (INTRO) in both languages of our bilingual children, who also demonstrated over-reliance on overt-marking of definiteness in Mandarin. The results corroborate previous studies on children's referential abilities, suggesting that linguistic properties involving morphosyntactic structure, information structure, and discourse could be particularly vulnerable in bilingual grammars. Regarding the role of input, our results underscore the importance of structural frequency in the input in shaping patterns of bilingual reference production. We have discovered mother–child association in the production of felicitous REs, the strength of which was modulated by working memory across language and discourse function. Amount of exposure did not seem to predict referential choice in our bilingual children. We postulated that there might be thresholds for amount of exposure to influence reference production. These findings shed lights on how language, input and cognitive skills might jointly influence bilingual reference production. They have direct relevance and precise implications for practice. To boost the acquisition of REs, which involve syntax-semantics-discourse interfaces, increasing the amount of input would not work best for bilingual children with relatively high proficiency in both languages, and bilingual children with different working memory capacities may benefit from different pedagogical strategies tailored for them. Those equipped with better working memory may display immediate benefit from increased frequency of REs in the input, and those with weaker working memory may need supplementary training on working memory to show similar progress, presumably not only in RE but in language learning in general.

The findings of this small-scale exploratory study await replications with a larger sample of children with different language combinations and an array of language dominance profiles. Further investigations may tease apart cross-linguistic influence effect and study the threshold effect of amount of language exposure with a more focused design using multiple linguistic and cognitive measures (e.g., combining both performance-based tests and caregiver ratings).

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 Survey and Behavioural Research Ethics Committee of the Chinese University of Hong Kong. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

VY, ZM, and JZ designed the study and established collaboration with the kindergarten. JZ prepared the experimental materials and procedures, and revised and improved on the Zoom-based data collection protocols provided by ZM, with QC and YL's assistance. JZ, QC, and YL recruited the participants and collected the data. QC and YL transcribed and coded most of the data under the supervision of JZ. JZ analyzed and interpreted the data in consultation with ZM and VY. JZ wrote the first and second drafts of the paper. ZM revised both. All authors worked on refining the text. All authors contributed to the article and approved the submitted version.

Funding

We would like to acknowledge funding support from the project "Development of Bilingualism in Chinese-speaking and Overseas Communities" conducted by our three labs: the University of Cambridge – Chinese University of Hong Kong Joint Laboratory for Bilingualism, the Chinese University of Hong Kong – Peking University – University System of Taiwan Joint Research Centre for Language and Human Complexity, and the Childhood Bilingualism Research Centre at The Chinese University of Hong Kong. The work described in this paper was also partially supported by a General Research Fund from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 14615820), awarded to ZM.

Acknowledgments

We thank members of the Childhood Bilingualism Research Centre at the Chinese University of Hong Kong for their

unwavering support. Special thanks go to Ziyan Meng, Sihui Liu, Ruozhu Zou, Miriam Wong, Hecheng Zhang, Wenchun Yang, Angel Chan, Anna Ma, Shiyu He, and Elaine Lau for their help and suggestions for this project. We are grateful to the school boards, principals, teachers, and families who made this study possible. We would also like to thank our reviewers for their helpful suggestions and valuable feedback on this manuscript.

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.897031/full#supplementary-material

References

Aalberse, S., Zou, Y., and Andringa, S. (2017). "Extended use of demonstrative pronouns in two generations of mandarin Chinese speakers in the Netherlands. Evidence of convergence?" in *Cross-Linguistic Influence in Bilingualism: In Honor of Aafke Hulk*. eds. E. Blom, L. Cornips and J. Schaeffer (Amsterdam: John Benjamins), 25–48.

Allen, S., Skarabela, B., and Hughes, M. (2008). "Using corpora to examine discourse effects in syntax" in *Corpora in Language Acquisition Research: Finding Structure in Data*. ed. H. Behrens (Amsterdam: John Benjamins), 99–137.

Ambridge, B., Kidd, E., Rowland, C., and Theakston, A. (2015). The ubiquity of frequency effects in first language acquisition. *J. Child Lang.* 42, 239–273. doi: 10.1017/S030500091400049X

Archibald, L. M. D. (2017). Working memory and language learning: a review. *Child Lang. Teach. Ther.* 33, 5–17. doi: 10.1177/0265659016654206

Ariel, M. (1990). *Accessing Noun-Phrase Antecedents*. London: Routledge.

Bates, D., Maechler, 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

Belletti, A., Bennati, E., and Sorace, A. (2007). Theoretical and developmental issues in the syntax of subjects: evidence from near-native Italian. *Nat. Lang. Linguist. Theor.* 25, 657–689. doi: 10.1007/s11049-007-9026-9

Blom, E., Küntay, A., Messer, M., Verhagen, J., and Leseman, P. (2014). The benefits of being bilingual: working memory in bilingual Turkish–Dutch children. *J. Exp. Child Psychol.* 128, 105–119. doi: 10.1016/j.jecp.2014.06.007

Bock, K., Irwin, D. E., and Davidson, D. J. (2004). "Putting first things first" in *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. eds. J. Henderson and F. Ferreira (New York: Psychology Press), 249–278.

Chen, P. (2004). Identifiability and definiteness in Chinese. *Linguistics* 42, 1129–1184. doi: 10.1515/ling.2004.42.6.1129

Chen, Y., Lee, P., and Pan, H. (2016). "Topic and focus marking in Chinese" in *The Oxford Handbook of Information Structure*. eds. C. Féry and S. Ishihara (Oxford: Oxford University Press), 733–752.

Chen, L., and Lei, J. (2012). The production of referring expressions in oral narratives of Chinese-English bilingual speakers and monolingual peers. *Child Lang. Teach. Ther.* 29, 41–55. doi: 10.1177/0265659012459527

Chen, J., and Narasimhan, B. (2018). "Information structure and ordering preferences in child and adult speech in English," in The Proceedings of the 42nd Boston University Conference on language development. eds. A. B. Bertolini and M. J. Kaplan (Boston: Cascadilla Press), 131–139.

Chen, J., Narasimhan, B., Chan, A., Yang, W., and Yang, S. (2020). Information structure and word order preference in child and adult speech of mandarin Chinese. *Language* 5:14. doi: 10.3390/languages5020014

Cheng, L. L.-S., and Sybesma, R. (1999). Bare and not-so-bare nouns and the structure of NP. *Linguist. Inquiry* 30, 509–542. doi: 10.1162/002438999554192

Colozzo, P., and Whitely, C. (2014). Keeping track of characters: factors affecting referential adequacy in children's narratives. *First Lang.* 34, 155–177. doi: 10.1177/0142723714522164

De Cat, C. (2015). "The cognitive underpinnings of referential abilities" in *The Acquisition of Reference (Trends in Language Acquisition Research, Vol. 15)*. eds. L. Serratrice and S. E. Allen (Amsterdam: John Benjamins), 263–283.

Dunn, M., and Dunn, L. M. (2007). *Peabody Picture Vocabulary Test-4*. Circle Pines, MN: AGS.

Ebert, K. D. (2014). Nonlinguistic cognitive effects of language treatment for children with primary language impairment. *Commun. Disord. Q.* 35, 216–225. doi: 10.1177/1525740114523311

Espy, K. A., Sheffield, T. D., Wiebe, S. A., Clark, C. A. C., and Moehr, M. J. (2011). Executive control and dimensions of problem behaviors in preschool children. *J. Child Psychol. Psychiatry* 52, 33–46. doi: 10.1111/j.1469-7610.2010.02265.x

Fernald, A. (2006). "When infants hear two languages: interpreting research on early speech perception by bilingual children," in *Childhood bilingualism: Research on infancy through school age*. eds. P. McCardle and E. Hoff (Bristol, Blue Ridge Summit: Multilingual Matters), 19–29. doi: 10.21832/9781853598715-003

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., et al. (2012). MAIN: multilingual assessment instrument for narratives. *ZAS Pap. Linguist.* 56, 1–140. doi: 10.21248/zaspil.56.2019.414

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., et al. (2015). Assessment of narrative abilities in bilingual children in *Assessing Multilingual Children Disentangling Bilingualism From Language Impairment*. eds. S. Armon-Lotem, J. de Jong and N. Meir (Bristol, Blue Ridge Summit: Multilingual Matters), 243–269.

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Bohnacker, U., et al. (2019). MAIN: multilingual assessment instrument for narratives. *ZAS Pap. Linguist.* 63, 1–36. doi: 10.21248/zaspil.63.2019.516

Garon, N. M., Piccinin, C., and Smith, I. M. (2016). Does the BRIEF-P predict specific executive function components in preschoolers? *Appl. Neuropsychol. Child* 5, 110–118. doi: 10.1080/21622965.2014.1002923

Gioia, G. A., Espy, K. A., and Isquith, P. K. (2003). *The Behavior Rating Inventory of Executive Function-Preschool Version (BRIEF-P)*. Odessa, FL: Psychological Assessment Resources

Green, P., and MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* 7, 493–498. doi: 10.1111/2041-210X.12504

Grüter, T., and Paradis, J. Eds. (2014). *Input and Experience in Bilingual Development (Trends in Language Acquisition Research, Vol. 13)*. Amsterdam: John Benjamins Publishing.

Gundel, J., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language* 69, 274–307. doi: 10.2307/416535

Hamdani, S., Kan, R., Chan, A., and Gagarina, N. (2021). Summarizing experience: identifying bilingual DLD using online testing. Invited talk given at the conference on "online elicitation of narrative texts: summarizing experience and making plans". 25-27 Jan 2021.

Hendriks, P. (2016). Cognitive modeling of individual variation in reference production and comprehension. *Front. Psychol.* 7:506. doi: 10.3389/fpsyg.2016.00506

Hickmann, M., and Hendriks, H. (1999). Cohesion and anaphora in children's narratives: a comparison of English, French, German, and Chinese. *J. Child Lang.* 26, 419–452. doi: 10.1017/S0305000999003785

Hickmann, M., Hendriks, H., Roland, F., and Liang, J. (1996). The marking of new information in children's narratives: a comparison of English, French, German and mandarin Chinese. *J. Child Lang.* 23, 591–619. doi: 10.1017/S0305000900008965

Hickmann, M., and Liang, J. (1990). Clause-structure variation in Chinese narrative discourse: a developmental analysis. *Linguistics* 28, 1167–1200. doi: 10.1515/ling.1990.28.6.1167

Hickmann, M., Schimke, S., and Colonna, S. (2015). "From early to late mastery of reference: multifunctionality and linguistic diversity" in *The Acquisition of Reference (Trends in Language Acquisition Research, Vol. 15)*. eds. L. Serratrice and S. E. Allen (Amsterdam: John Benjamins), 181–211.

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

Hoff, E., Core, C., and Shanks, K. F. (2020). The quality of child-directed speech depends on the speaker's language proficiency. *J. Child Lang.* 47, 132–145. doi: 10.1017/S030500091900028X

Hoff, E., Welsh, S., Place, S., and Ribot, K. M. (2014). "Properties of dual language input that shape bilingual development and properties of environments that shape dual language input," in *Input and experience in bilingual development (trends in language acquisition research, Vol. 13)*. eds. T. Grüter and J. Paradis (Amsterdam: John Benjamins Publishing), 119–140.

Hulk, A., and Müller, N. (2000). Bilingual first language acquisition at the interface between syntax and pragmatics. *Biling. Lang. Congn.* 3, 227–244. doi: 10.1017/S1366728900000353

Indrarathne, B., and Kormos, J. (2018). The role of working memory in processing L2 input: insights from eye-tracking. *Biling. Lang. Congn.* 21, 355–374. doi: 10.1017/S1366728917000098

Jia, R., and Paradis, J. (2015). The use of referring expressions in narratives by mandarin heritage language children and the role of language environment factors in predicting individual differences. *Biling. Lang. Congn.* 18, 737–752. doi: 10.1017/S1366728914000728

Kail, M., and Hickmann, M. (1992). French children's ability to introduce referents in narratives as a function of mutual knowledge. *First Lang.* 12, 73–94. doi: 10.1177/014272379201203405

Kronenberger, W. G., Montgomery, C. J., Henning, S. C., Ditmars, A., Johnson, C. A., Herbert, C. J., et al. (2021). Remote assessment of verbal memory in youth with Cochlear implants during the COVID-19 pandemic. *Am. J. Speech Lang. Pathol.* 30, 740–747. doi: 10.1044/2021_AJSLP-20-00276

Li, C., and Thompson, S. (1981). *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Lin, S., Keysar, B., and Epley, N. (2010). Reflexively mindblind: using theory of mind to interpret behavior requires effortful attention. *J. Exp. Soc. Psychol.* 46, 551–556. doi: 10.1016/j.jesp.2009.12.019

Lindgren, J., Reichardt, V., and Bohnacker, U. (2020). Character introductions in oral narratives of Swedish–German bilingual preschoolers. *First Lang.* 42, 234–262. doi: 10.1177/0142723719897440

Luo, J., Yang, W. C., Chan, A., Cheng, K., Kan, R., and Gagarina, N. (2020). *The Multilingual Assessment Instrument for Narratives (MAIN): Adding Mandarin to MAIN. ZAS Papers in Linguistics (ZASPiL)*. Berlin: ZAS, 64.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk: Transcription Format and Programs. 3rd Edn*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Mai, Z., and Deng, X. (2019). Selective vulnerability and dominant language transfer in the acquisition of the Chinese cleft construction by heritage speakers. *Linguist. Approach. Bilingual.* 9, 202–227. doi: 10.1075/lab.16040.mai

Mai, Z., Zhao, L., and Yip, V. (2021). The mandarin Ba-construction in school-age heritage speakers and their parental input. *Linguist. Approach. Bilingual.* 12, 377–405. doi: 10.1075/lab.18025.mai

Martini, M., Sachse, P., Furtner, M. R., and Gaschler, R. (2015). Why should working memory be related to incidentally learned sequence structures. *Cortex* 64, 407–410. doi: 10.1016/j.cortex.2014.05.016

Min, R.-F. (1994). The acquisition of referring expressions by young Chinese children: A longitudinal study of the forms and functions of early noun phrases. Unpublished doctoral dissertation. Nijmegen, the Netherlands: Catholic University of Nijmegen.

Miyake, A., Friedman, N., Emerson, M., Witzki, A., Howerter, A., and Wager, T. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. *Cogn. Psychol.* 41, 49–100. doi: 10.1006/cogp.1999.0734

Morales, J., Calvo, A., and Bialystok, E. (2013). Working memory development in monolingual and bilingual children. *J. Exp. Child Psychol.* 114, 187–202. doi: 10.1016/j.jecp.2012.09.002

Moro, F. (2016). Dynamics of Ambon Malay: Comparing Ambon and the Netherlands. Unpublished doctoral dissertation. Radboud University, Nijmegen.

Narasimhan, B., and Dimroth, C. (2008). Word order and information status in child language. *Cognition* 107, 317–329. doi: 10.1016/j.cognition.2007.07.010

Nilsen, E. S., and Bacso, S. A. (2017). Cognitive and behavioural predictors of adolescents' communicative perspective-taking and social relationships. *J. Adolesc.* 56, 52–63. doi: 10.1016/j.adolescence.2017.01.004

Nilsen, E. S., and Graham, S. A. (2009). The relations between children's communicative perspective-taking and executive functioning. *Cogn. Psychol.* 58, 220–249. doi: 10.1016/j.cogpsych.2008.07.002

O'Meagher, S., Norris, K., and Kemp, N. (2018). Examining the relationship between performance-based and questionnaire assessments of executive function in young preterm children: implications for clinical practice. *Child Neuropsychol.* 25, 899–913. doi: 10.1080/09297049.2018.1531981

Otwinowska, A., Mieszkowska, K., Białecka-Pikul, M., Opacki, M., and Haman, E. (2020). Retelling a model story improves the narratives of polish-English bilingual children. *Int. J. Biling. Educ. Biling.* 23, 1083–1107. doi: 10.1080/13670050.2018.1434124

Paradis, J. (2017). Parent report data on input and experience reliably predict bilingual development and this is not trivial. *Biling. Lang. Congn.* 20, 27–28. doi: 10.1017/S136672891600033X

Paradis, J., and Navarro, S. (2003). Subject realization and crosslinguistic interference in the bilingual acquisition of Spanish and English: what is the role of the input? *J. Child Lang.* 30, 371–393. doi: 10.1017/S0305000903005609

Paradis, J., Tremblay, A., and Crago, M. (2014). "French-English bilingual children's sensitivity to child-level and language-level input factors in morphosyntactic acquisition" in *Input and Experience in Bilingual Development (Trends in Language Acquisition Research, Vol. 13)*. eds. T. Grüter and J. Paradis (Amsterdam: John Benjamins Publishing), 161–180.

Pearson, B. (2007). Social factors in childhood bilingualism in the United States. *Appl. Psycholinguist.* 28, 399–410. doi: 10.1017/S014271640707021X

Pennington, B. F., and Ozonoff, S. (1996). Executive functions and developmental psychopathology. *J. Child Psychol. Psychiatry Allied Discip.* 37, 51–87. doi: 10.1111/j.1469-7610.1996.tb01380.x

Place, S., and Hoff, E. (2016). Effects and noneffects of input in bilingual environments on dual language skills in 2 ½-year-olds. *Biling. Lang. Congn.* 19, 1023–1041. doi: 10.1017/S1366728915000322

Polinsky, M. (2006). Incomplete acquisition: American Russian. *J. Slavic Linguist.* 14, 191–262. Available at: https://www.jstor.org/stable/24599616

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/

Sah, W.-H. (2018). Referential choice in narratives of mandarin-speaking children with autism spectrum disorder: form, function, and adequacy. *First Lang.* 38, 225–242. doi: 10.1177/0142723717739198

Serratrice, L. (2007). Referential cohesion in the narratives of bilingual English–Italian children and monolingual peers. *J. Pragmat.* 39, 1058–1087. doi: 10.1016/j.pragma.2006.10.001

Serratrice, L., and De Cat, C. (2020). Individual differences in the production of referential expressions: the effect of language proficiency, language exposure and executive function in bilingual and monolingual children. *Biling. Lang. Congn.* 23, 371–386. doi: 10.1017/S1366728918000962

Sheng, L., Shi, H., Wang, D., Hao, Y., and Zheng, L. (2020). Narrative production in mandarin-speaking children: effects of language ability and elicitation method. *J. Speech Lang. Hear. Res.* 63, 774–792. doi: 10.1044/2019_JSLHR-19-00087

Singapore Department of Statistics (2020). Census of population 2020 statistical release 1: Demographic characteristics, education, language and religion. Available at: https://www.singstat.gov.sg/-/media/files/publications/cop2020/sr1/cop2020sr1.pdf

Sorace, A., and Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Second. Lang. Res.* 22, 339–368. doi: 10.1191/0267658306sr271oa

Sorace, A., Serratrice, L., Filiaci, F., and Baldo, M. (2009). Discourse conditions on subject pronoun realization: testing the linguistic intuitions of older bilingual children. *Lingua* 119, 460–477. doi: 10.1016/j.lingua.2008.09.008

Sunderman, G., and Kroll, J. F. (2009). When study-abroad experience fails to deliver: the internal resources threshold effect. *Appl. Psycholinguist.* 30, 79–99. doi: 10.1017/S0142716408090048

Sybesma, R. P. E. (1992). *Causatives and Accomplishments: The Case of Chinese Ba*, Doctoral dissertation, Holland Institute of Generative Linguistics (HIL)/Leiden University.

Thordardottir, E. (2014). "The typical development of simultaneous bilinguals" in *Input and Experience in Bilingual Development (Trends in Language Acquisition research, Vol. 13)*. eds. T. Grüter and J. Paradis (Amsterdam: John Benjamins Publishing), 141–160.

Torregrossa, J. (2017). "The role of executive functions in the acquisition of reference: the production of demonstrative pronouns by German monolingual children," in *Language Acquisition at the Interfaces: Proceedings of GALA (Generative Approaches to Language Acquisition)*. eds. J. Choi, H. Dimirdache, O. Lungu and L. Voeltzel (Newcastle upon Tyne: Cambridge Scholars Publishing), 318–331.

Torregrossa, J., Andreou, M., Bongartz, C., and Tsimpli, I. (2021). Bilingual acquisition of reference: the role of language experience, executive functions and cross-linguistic effects. *Biling. Lang. Congn.* 24, 694–706. doi: 10.1017/S1366728920000826

Unsworth, S. (2013). Assessing the role of current and cumulative exposure in simultaneous bilingual acquisition: the case of Dutch gender. *Biling. Lang. Congn.* 16, 86–110. doi: 10.1017/S1366728912000284

Unsworth, S. (2014). "Comparing the role of input in bilingual acquisition across domains," *Input and Experience in Bilingual Development (Trends in Language Acquisition Research, Vol. 13)*. eds. T. Grüter and J. Paradis (Amsterdam: John Benjamins Publishing), 181–201.

van Kleeck, A., Gillam, R. B., and Hoffman, L. M. (2006). Training in phonological awareness generalizes to phonological working memory: a preliminary investigation. *J. Speech Lang. Pathol.* 1, 228–243. doi: 10.1037/h0100201

van Rij, J. (2012). Pronoun processing: Computational, behavioral, and psychophysiological studies in children and adults. Ph.D. dissertation. University of Groningen, Groningen.

Wardlow, L., and Heyman, G. D. (2016). The roles of feedback and working memory in children's reference production. *J. Exp. Child Psychol.* 150, 180–193. doi: 10.1016/j.jecp.2016.05.016

Wechsler, D. (2013). *Wechsler Preschool and Primary Scale of Intelligence Taiwan manual. 4th Edn.* (Mandarin Chinese revised version by C. Chen, R. Chen et al.). Taipei: Chinese Behavioral Science Corporation.

Werfel, K. L., Grey, B., Johnson, M., Brooks, M., Cooper, E., Reynolds, G., et al. (2021). Transitioning speech-language assessment to a virtual environment: lessons learned from the ELLA study. *Lang. Speech Hear. Serv. Sch.* 52, 769–775. doi: 10.1044/2021_LSHSS-20-00149

Wong, A. M. Y., and Johnston, J. R. (2004). The development of discourse referencing in Cantonese speaking children. *J. Child Lang.* 31, 633–660. doi: 10.1017/s030500090400604x

Wu, Z., Huang, R., and Zhang, Z. (2015). Hanyu (bu) dingzhi biaoji ertong xide yanjiu. *Foreign Lang. Teach. Res.* 47, 176–189.

Zhou, J., Mai, Z., and Yip, V. (2021). Bidirectional cross-linguistic influence in object realization in Cantonese-English bilingual children. *Biling. Lang. Congn.* 24, 96–110. doi: 10.1017/S1366728920000231

Zhou, J., and Yip, V. (2021). The post-verbal pronoun KEOI in child Cantonese: a corpus-based study. *J. Chin. Linguist.* 49, 379–419. doi: 10.1353/jcl.2021.0012

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 10 October 2022
DOI 10.3389/fcomm.2022.841543

OPEN ACCESS

EDITED BY
Angel Chan,
Hong Kong Polytechnic University,
Hong Kong SAR, China

REVIEWED BY
Eliseo Diez-Itza,
University of Oviedo, Spain
Lindsay Butler,
Boston University, United States

*CORRESPONDENCE
Elisa Mattiauda
elisa.mattiauda.18@ucl.ac.uk

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 22 December 2021
ACCEPTED 12 September 2022
PUBLISHED 10 October 2022

CITATION
Mattiauda E, Hassiotis A and Perovic A (2022) Narrative language abilities in adults with Down syndrome: A remote online elicitation study using the Multilingual Assessment Instrument for Narratives (MAIN). *Front. Commun.* 7:841543. doi: 10.3389/fcomm.2022.841543

Narrative language abilities in adults with Down syndrome: A remote online elicitation study using the Multilingual Assessment Instrument for Narratives (MAIN)

Elisa Mattiauda[1]*, Angela Hassiotis[2] and Alexandra Perovic[1]

[1]Division of Psychology and Language Sciences, University College London (UCL), London, United Kingdom, [2]Division of Psychiatry, University College London (UCL), London, United Kingdom

Introduction: This research represents, to the best of our knowledge, the first attempt at assessing narrative retell remotely in people with Down syndrome and will provide valuable information on the validity and feasibility of remote online assessment with this population. Most research on language abilities in Down syndrome has focused on children and adolescents, making adults an understudied population. The present research seeks to establish a baseline of functioning for narrative language abilities in adults with Down syndrome, as part of a larger research aiming to investigate possible changes associated with aging and the emergence of Alzheimer's disease in this population.

Methods: We recruited 13 adolescents and young adults with Down syndrome aged 15–33 years (mean age: 21), matched to a control group of younger typically developing children aged 4–10 years (mean age: 6) on verbal Mental Age (MA). Participants completed a picture-based story retell activity from the Multilingual Assessment Instrument for Narratives (MAIN) and a series of standardized background measures of language and cognitive ability.

Results: Our analyses focused on macrostructural indices of narrative performance, narrative length and lexical diversity. Results revealed that our participants with Down syndrome were outperformed by verbal MA-matched controls on measures of story structure and story comprehension, as well as lexical diversity. No difference was found on total number of words, indicating the groups produced comparable amounts of speech despite differences in story grammar and lexis.

Discussion: We interpret the results in light of previous research on macrostructural narrative performance in adults and younger adolescents with Down syndrome. Recruitment and data collection outcomes are discussed in terms of successful strategies and possible improvements. We conclude that remote online assessment of people with Down syndrome is feasible, although considerations should be made with regards to facilitating enrolment, and task engagement.

Our participants demonstrated ability to engage with the experimenters over video chat and were able to complete the activities proposed mostly independently, with minimal involvement required from caregivers. Recommendations for future remote online studies involving children and people with intellectual disabilities are discussed.

KEYWORDS

Down syndrome, intellectual disability, narrative language, online remote elicitation, macrostructure

Introduction

Down syndrome is a developmental disorder associated with mild to moderate levels of intellectual disability (Chapman and Hesketh, 2000). In 95% of cases, the syndrome is caused by a full extra copy of chromosome 21 (trisomy), while in a small proportion of cases the extra copy is only present in some cells (mosaicism) or parts of chromosome 21 attach to another chromosome (translocation) (Martin et al., 2009). Down syndrome affects around 1 in 1,000 live births (Wu and Morris, 2013) and according to a 2015 estimate (de Graaf et al., 2021) over 41,000 people with Down syndrome live in the United Kingdom alone.

Over the past 2 years, the population with Down syndrome, amongst others, has been disproportionally affected by changes associated with the Covid-19 pandemic, particularly by the strict social-isolation regulations that have been enforced worldwide. Individuals with Down syndrome are prone to a range of physical and psychiatric complications associated with the syndrome which make them particularly vulnerable to the Covid-19 disease (Clift et al., 2021). In addition to more severe health risks associated with the SARS-CoV-2 virus, recent reports on the impact of the pandemic have evidenced an increase in depressive symptoms and worse overall functioning in adults with Down syndrome as a result of local lockdown measures (Villani et al., 2020). In light of the increased health risks associated with the current climate, research practices worldwide have had to adapt their methodology to minimize face-to-face contact with participants. As a result, many research teams have turned to remote approaches to language assessment. In the following sections, we review the cognitive and language profile of adults with Down syndrome, with a focus on aspects of narrative language abilities, and present the rationale for the present research. The study seeks to bring insights into the narrative language skills of adolescents and adults with Down syndrome and assess the feasibility of a remote online approach to recruitment and language elicitation in this population.

Cognitive and language abilities in Down syndrome

While a certain degree of heterogeneity must be recognized, individuals with Down syndrome often present with overall developmental delays, coupled with selective weaknesses in aspects of higher-order cognitive functioning. General patterns include later onset of language acquisition, impaired speech production, and impairments in working memory, especially phonological as opposed to visuospatial memory (Jarrold and Baddeley, 2001; Campbell et al., 2013). Both memory and language exhibit significant delays in this population relative to typically developing (TD) counterparts, however abilities in these domains are far from being homogeneously affected. Language, in particular, presents intriguing dissociations. Receptive language skills tend to be relatively stronger than production (Chapman et al., 1998; Miles and Chapman, 2002; Cleave et al., 2012), with vocabulary comprehension generally aligning with non-verbal mental age expectations (Abbeduto et al., 2003). Expressive language, on the other hand, is an area of marked difficulty: most individuals with Down syndrome produce shorter and more simplified sentences relative to what would be expected on the basis of non-verbal mental age (Chapman et al., 1998; Caselli et al., 2008; Price et al., 2008). Morphosyntax is particularly weak. Difficulties are reported in both comprehension and production of grammatical morphemes, with common errors involving omission or incorrect use of past tense -ed, third person singular -s, present progressive -ing, auxiliaries and articles (Fowler et al., 1994; Hesketh and Chapman, 1998; Eadie et al., 2002; Caselli et al., 2008). At the level of syntax, individuals with Down syndrome also show difficulties in the comprehension of complex sentences, such as passives, relative clauses, and interrogatives (e.g., Joffe and Varlokosta, 2007; Frizelle et al., 2019; Perovic and Wexler, 2019), as well as specific syntactic relations involved in the interpretation of reflexive pronouns (Perovic, 2006), across their lifespan. Less is known about the pragmatic skills of individuals with Down syndrome. Early studies report pragmatics to be a relative strength for both

children and adults with Down syndrome, especially compared to their grammatical skills (see Roberts et al., 2007, for a review). However, more recent investigations suggest that children with Down syndrome may perform poorer than younger controls in almost all areas of pragmatics, from topic initiation, topic elaboration and maintenance, to the use of context and conversational repairs (e.g., Smith et al., 2017).

Most of the available evidence on language abilities in Down syndrome comes from studies carried out with children and adolescents, making adults an understudied population. The relatively recent increase in life expectancy for people with Down syndrome (Strauss and Eyman, 1996) has invited interest into the cognitive changes that may be associated with aging in this population. Clinical studies have established a now well-researched association between Alzheimer's disease (AD) and the syndrome, indicating that people with Down syndrome are at ultra-high risk of developing the neurodegenerative condition (Sinai et al., 2012; Startin et al., 2019a). It is estimated that roughly 75% of individuals with Down syndrome aged over the age of 60 show clinical markers of Alzheimer's dementia (Lai and Williams, 1989, also see McCarron et al., 2014 and McCarron et al., 2017 for similar estimates), though clinical changes can be detected as early as 35 years of age (see Ballard et al., 2016). However, lack of a clear understanding of AD symptom progression in Down syndrome (see Lautarescu et al., 2017 for a review) brings the issue of (early) diagnosis in this population front and center. In fact, early detection of neurodegenerative conditions in populations with intellectual disability is greatly complicated by the presence of pre-existing lifelong cognitive impairments which make diagnosis difficult in the absence of previous baseline assessments (Devenny et al., 2000; Sinai et al., 2012). Few studies have focused on the extent to which language skills are affected by age- and dementia-related decline in the Down syndrome population, revealing inconsistent findings. For example, Devenny and Krinsky-McHale (1998) report no evidence of age-dependent language deterioration in adults with Down syndrome, while others have supported the idea of a decline in language skills from the fourth decade of life (e.g., Carter Young and Kramer, 1991; Cooper and Collacott, 1995; Perovic and Wexler, 2019). Nevertheless, a more recent publication argues that the inclusion of language assessment can be beneficial to improving the diagnosis of dementia in Down syndrome (Pulsifer et al., 2020). This suggests that monitoring language skills in adults with Down syndrome could generate useful insights into age- and AD-related changes in this population. In particular, honing in on language domains that could reveal language difficulties resulting from Alzheimer-related decline in adults without Down syndrome, such as narrative skills (e.g., Chapman et al., 1998; Ash et al., 2007), would provide valuable relevant information. Such information, in turn, would allow us to precisely map out the linguistic profile of the adult population with Down syndrome, and pinpoint the areas of language ability that could be problematic for those adults with Down syndrome at risk of developing Alzheimer's.

Narrative abilities in Down syndrome

One particularly fruitful method of assessing language production abilities of individuals with Down syndrome at both syntactic and pragmatic level relies on narrative tasks. Narratives are complex forms of discourse consisting of connected passages that primarily integrate information about events ("landscape of action") and participants' mental states ("landscape of consciousness," Bruner, 1986). Such productions, therefore, require careful coordination of interacting socio-cognitive and linguistic factors. On the one hand, well-structured narratives construct a hierarchical framework for the presentation of related events, while incorporating different levels of perspective from both the narrator and the participants in the story (Pearson and De Villiers, 2006). As such, storytelling ability is influenced by the speaker's prior life experiences and world knowledge, including their familiarity with themes and events in the story, their understanding of event sequences, temporal or causal relationships, and more general social and cultural dynamics (Miles and Chapman, 2002; Segal and Pesco, 2015). Being able to reason about others' mental states (the ability traditionally referred to as Theory of Mind, ToM) represents an important aspect of narrative competence, as it allows the narrator to make predictions and offer interpretations of characters' behaviors, as well as judgements about the listener's perspective and the common ground shared with the audience—all necessary to produce informative and relevant narratives (Matthews et al., 2018).

On the other hand, the expression of aspects relevant to event structure and participant's states will be influenced by the underlying linguistic competence of the speaker and their ability to formulate, through means of lexical choices and morphosyntactic organization, the linguistic scaffolding necessary for generating cohesive and comprehensible narratives. Evidence suggests, in fact, that pragmatic language skills such as the ones involved in storytelling are in large part directly related to formal language competence (Matthews et al., 2018), a finding which could have important consequences on how we interpret narrative performance. For example, a lack of references to character's mental states in narrative production could be indicative of ToM difficulties, but could also be explained in terms of an underlying syntactic impairment: mental state verbs often require more complex syntactic constructions (e.g., complement clauses) to be expressed, which could reveal problematic for young children and especially for individuals with language impairments (de Villiers and de Villiers, 2000 for an account of the role of language in the development of ToM). Production of narrative content, especially for individuals with intellectual disabilities, may

also be affected by sampling context: it has been evidenced that participants with DS tend to produce larger MLUs and more complex syntactic constructions in response to picture-supported narrative elicitation techniques as opposed to conversational samples, including pictureless narration (Miles and Sindberg, 2006). Narrative tasks can thus provide useful insights into a range of underlying linguistic, cognitive and social abilities of speakers, and have proven especially effective modes of language elicitation in populations with language and intellectual impairments (Norbury and Bishop, 2003). Particularly, in addition to providing a structured framework for the elicitation of speech samples (Sealey and Gilmore, 2008), narratives offer a way of evaluating production on two distinct levels of linguistic competence: macrostructure and microstructure. Macrostructure refers to higher order organizational and cohesive aspects of story structure, while microstructure refers to internal elements of linguistic constructions. Macrostructural analysis focuses on narrative competence from the perspective of hierarchical organization, by examining the presentation of story grammar elements and the structural complexity of story episodes. This might involve determining the number and structure of episodes contained in a narrative by identifying key elements such as initiating events, goals, and outcomes, as well as evaluating the use of narrative tools such as appendages, orientations and evaluations (Berman and Slobin, 1994). On the other hand, microstructural analysis focuses on evaluating the use and accuracy of morphosyntactic constructions, such as verb morphology, syntactic phrases and dependent clauses (Ukrainetz et al., 2005; Justice et al., 2006). In our brief review of studies examining narratives in adults with Down syndrome, we shall focus on macrostructure rather than microstructure. Microstructural skills are expected to be deficient across the lifespan, in view of the persistent grammatical deficits associated with Down syndrome. Macrostructural abilities, however, may show different patterns. As discussed above, pragmatics is generally considered an area of relative strength in individuals with DS. Since discourse cohesion and coherence are domains known to be affected by age-related decline in the population without Down syndrome (e.g., Ash et al., 2007), however, aspects of macrostructure may prove to be even more vulnerable in adults with Down syndrome who are approaching the age of dementia onset for this population.

Below we review a selection of studies that have examined macrostructural narrative language skills in individuals with Down syndrome. In light of the limited availability of studies retrieved focusing on *adult* language skills, we also include research carried out with older children and adolescents with Down syndrome. In line with previously discussed considerations regarding the effects of task structure and sampling context on the elicitation of narrative samples that can be compared across populations (Sealey and Gilmore, 2008), we review studies that investigated structured narratives, similar to ours, elicited with the support of pictorial stimuli.

In a study that elicited narratives relying on the wordless picture book, "Frog where are you?", from a group of 33 English-speaking children and young adults with Down syndrome aged between 12 and 26 years (mean age: 18.76), Miles and Chapman (2002) reported that expression of plot line was commensurate with non-verbal mental age and syntax comprehension. According to the report, thematic content was conveyed at a level consistent with their syntax comprehension as opposed to expressive language. However, when compared to TD controls matched on a measure of expressive language such as mean length of utterance (MLU), participants with Down syndrome expressed significantly more plot line events, thematic content and episodic events relating to character misadventures. Overall, the authors conclude that the narratives of participants with Down syndrome in this study indicate levels of conceptual abilities that are in line with syntax comprehension and non-verbal mental age, but tend to exceed expressive language ability. It appears, then, that macrostructural elements of narrative skill may not be accurately predicted by measures of expressive language such as MLU alone.

Finestack et al. (2012) examined the macrostructural story-telling skills of 24 English-speaking children and young adults with Down syndrome, aged between 12 and 23 years (mean age 16;9), compared to TD controls and individuals with Fragile X. Narratives were elicited using the wordless picture book "Frog goes to dinner." Participants with Down syndrome outperformed the non-verbal mental age-matched TD group on all macrostructural story components produced (Character Development, Mental States, Referencing, Conflict/Resolution, Cohesion, and Conclusion), with the exception of the Introduction component. When confronted with an MLU-matched control group, however, participants with Down syndrome did not show an advantage on any macrostructural dimensions.

In their investigation of a sample of participants that overlaps with that of Finestack et al. (2012) above, Channell et al. (2015) evaluated narrative abilities in 23 adolescents with Down syndrome aged between 10 and 16 years (mean age: 12). Participants with Down syndrome were reported to produce fewer episodic story elements compared to TD children matched for non-verbal cognitive ability. However, such difference was removed once MLU was controlled for, leading the authors to advance that individuals with Down syndrome do possess the necessary conceptual knowledge to express event-based story elements at a level consistent with non-verbal reasoning ability, but are limited in their expression of such elements by under-developed syntactic skills.

Zanchi et al. (2021) used an 18-picture storybook created to elicit the narratives of 13 Italian-speaking children and adolescents with Down syndrome, aged between 10 and 16 years (mean age: 12). No differences were observed on the

macrostructural elements in the stories of participants with Down syndrome and two groups of control participants, one matched on non-verbal MA (aged 5;2) and the other on MLU (5;5): all participants produced stories with comparable narrative structure and amount of event-based information.

The above studies typically control for MLU, or use MLU to match individuals with Down syndrome to control participants who are typically developing. However, while MLU is a useful measure of grammatical development in young TD children, its validity has been questioned in comparisons of grammatical mastery of children older than 4, both typically developing and in language-impaired populations (Scarborough et al., 1991). It is well-known that older children and young adults with Down syndrome will have richer vocabularies, producing more words in their utterances, but may still lack inflectional morphemes marking grammatical contrasts such as tense, which TD children acquire by about the age 4 (control participants in the studies above are aged 4 or 5). Studies relying on matching participants on vocabulary measures may thus provide richer insights into the narrative abilities of individuals with Down syndrome. The only study that focused solely on adults with Down syndrome and used vocabulary as a matching measure, Martzoukou et al. (2020), examined the story retelling abilities of 20 Greek-speaking adults with Down syndrome aged between 19 and 46 years (mean age: 28;2) by using two story retelling activities from the LITMUS-MAIN tool (Gagarina et al., 2012). The use of MAIN is of particular interest, as the task presents with several strengths. Firstly, MAIN offers a valuable tool for the controlled elicitation of narrative samples, while providing a detailed framework for the evaluation of macrostructural components of narrative ability. Furthermore, MAIN constitutes a particularly adept tool for the elicitation of expressive language in populations affected by developmental delays, as the picture-based retell activity in particular helps reduce cognitive load during narrative production (compared to those that use wordless picture books without providing a model story; Sealey and Gilmore, 2008). Secondly, since its development, the tool has been adapted to a variety of languages and cultural contexts, and has been adopted in a plethora of studies carried out all over the world to assess language abilities in typical and atypical populations (see https://main.leibniz-zas.de/en/worldwide-network/). As such, the task offers unique opportunities for cross-study comparisons and replications across a variety of populations. In their investigation of macrostructure, Martzoukou et al. (2020) report that individuals with Down syndrome performed worse on story structure relative to expressive vocabulary-matched TD controls, but no different to TD controls matched on non-verbal mental age. Comparison of participants' use of terms describing characters' internal states and emotions revealed that adults with Down syndrome used fewer such terms compared to both control groups. In addition, individuals with Down syndrome were found to perform similarly to non-verbal mental age matched controls and significantly worse than expressive vocabulary-matched controls on a series of comprehension questions tapping into the internal states and goals of characters in the story.

There are currently no studies reporting results from a remote collection of narratives from individuals with DS. Nonetheless, recent unpublished data suggest that this mode of elicitation of narratives is successful with young TD children (Sultana, 2022) and children with other developmental disorders such as autism (El-Raziq, 2022). Studies collecting natural language samples remotely provide reassurance that the quality of data obtained through parent-child interaction *via* video chat is of comparable quality to that obtained in the more traditional lab setting, for both TD children (Manning et al., 2020), and children with autism (Butler et al., 2022). Importantly, no differences in the crucial characteristics of TD children's samples such as speech intelligibility, lexical variety, grammatical errors or MLU were reported in samples obtained in person vs. video chat in Manning et al. (2020). Though such data are not yet available for individuals with DS, Kelleher et al. (2020) suggest that it is feasible to expect high quality data from unstructured parent–child interaction for infants with DS.

The current study

The present study examines the narrative language abilities of adolescents and young adults with Down syndrome by comparing them to a sample of TD controls matched on age of vocabulary comprehension. This represents part of a broader research project seeking to characterize the language skills of adults with Down syndrome and examine changes in language skills occurring over the course of adult life, particularly in relation to cognition and chronological age. With the current paper, we aim to establish a baseline for the general level of narrative language functioning achieved in early adulthood, by reporting on the abilities of a sample of young adults assessed before any suspected decline may have taken place. In the age range investigated here (15–35 years), cognition has not been affected by decline and language development is deemed to be approaching adult-like levels of syntactic organization and pragmatic proficiency. We expect this attainment to be reflected in our sample's story re-telling abilities. To elicit structured and comparable speech samples, we presented English-speaking participants with a story retelling activity from the Multilingual Assessment Instrument for Narratives (MAIN; Gagarina et al., 2019). In our study, the task was administered remotely during a video call with the participant and a caregiver. As such, in addition to being one of a handful of studies examining narrative language in adults with Down syndrome, the current study also represents the first attempt, to the best of our knowledge, at assessing narrative language skills remotely in a sample of people with Down syndrome. In this report we focus

on macrostructural aspects of narrative ability by examining performance on three dimensions of the MAIN retell task: Story Structure, production of Internal State Terms (IST), and Story Comprehension. We also analyse the narrative productions in terms of narrative length (total number of words produced) and lexical diversity (number of different words produced). The performance of our sample is compared to that of a control group of younger typically developing children matched on age of vocabulary comprehension. We asked the following questions about the abilities of our participant sample:

1. Will adolescents and young adults with Down syndrome produce narratives of length and lexical diversity comparable to younger vocabulary-matched TD controls?

2. Will adolescents and young adults with Down syndrome produce narratives with story structure, internal state terms and comprehension scores comparable to those of a younger TD control group matched on vocabulary comprehension?

3. Will the online approach be successful in eliciting narratives remotely and assessing verbal and non-verbal skills of individuals with DS?

In line with the previously reviewed literature on narrative skills in children and adolescents with Down syndrome, our participants should perform comparably well to matched TD controls on macrostructural properties of narrative retell. However, in line with the (sparse) literature on the narrative abilities of adults with this condition, it is also possible that our participants may perform less well-compared to verbal-MA matched controls, as reported in Martzoukou et al. (2020). Concerning the success of online elicitation with individuals with Down syndrome, while there is currently no literature on this topic, we expect that online assessment will be feasible based on recent literature with other populations with developmental disorders, such as autism.

Methods

Participants

Thirteen English-speaking participants with Down syndrome aged between 15;3 and 32;11 (M = 21.26) and 12 TD controls aged between 4;4 and 10;6 (M = 6.47 years) took part in the study. The chronological ages of TD children were matched to the verbal MAs of the participants with Down syndrome, as derived from the British Picture Vocabulary Scales-3 (BPVS-3) (see Ring and Clahsen, 2005, and Martzoukou et al., 2020, for similar matching methods).

Supplementary Table 1 shows demographic details of the two groups and descriptive background measures for the participants with Down syndrome: vocabulary comprehension (BPVS-3), grammar comprehension (TROG-2), non-verbal reasoning (KBIT-2) (see "Background measures" for more details).

All the participants with Down syndrome had a diagnosis of Trisomy 21. Two participants also had a diagnosis of ASD, while another had a diagnosis of obsessive-compulsive disorder (OCD) and bipolar disorder. None of the TD controls, all in primary school education, were reported to have any additional diagnosis or to be suspected of having any developmental delays.

As assessed by initial screening questions, six participants with Down syndrome had been diagnosed with hearing loss. Of these, four reported habitually using hearing aids. None reported other chronic illnesses or physical handicaps, and all reported speaking English as their main language. Two participants were reported to speak a second language non-fluently (Polish and Italian, respectively), while another three families reported speaking another language in the home (Konkani, Gujrati, French).

Four participants with Down syndrome were out of education at the time of testing, three attended secondary school and six were enrolled in further education. Participants with Down syndrome were reported to receive differing levels of support and attending various kinds of activities during the week. Some parents mentioned that their child received occupational and speech and language therapy (SLT) over the course of their lives, albeit irregularly, and no participant reported receiving regular ongoing SLT support. The lack of regular support at least for the past 2 years could be attributed partly to Covid-19, since many SLT services in the UK were significantly reduced during this time.

Participants with Down syndrome were recruited through online means. These included posts on social media platforms such as Facebook, and online adverts for the study shared by organizations such as the Down Syndrome Association (DSA). Though we did not collect specific information about where the participants heard about the study, Facebook posts seemed to be particularly fruitful when shared on groups dedicated to publicizing research opportunities for the Down syndrome community. A few of the participants were reached with the help of local charity organizations in London. Typical controls were recruited amongst contacts of the research team. Consent was provided by both participants and caregivers taking part. Participants with Down syndrome were provided with easy-read versions of the information and consent forms and TD controls completed child-appropriate versions of the form designed for different age groups. Most participants completed and returned the forms in digital version over email, while a small number requested a paper copy which was returned by post. Participant and caregiver were offered £10 vouchers each upon completion of the study.

The study was reviewed and approved by the ethical board of the Division of Psychology and Language Sciences at University College London (UCL).

Materials and procedure

The study involved a wide range of experimental measures and informant questionnaires which are not reported here. Below we provide the details and the administration procedure for both the LITMUS-MAIN narrative retell task and the three background measures that provide information about the general language and cognitive functioning of the participants in our sample.

LITMUS-MAIN narrative task

Narrative retell samples were elicited using the Cat story retelling activity from the revised version of the Multilingual Assessment Instrument for Narratives (MAIN; Gagarina et al., 2019), which was developed as part of the LITMUS test battery. In the current study, the task was adapted for online administration by following two previously developed adaptations: the first developed by Kapalková et al. (2021), for use with children, and another adaptation developed by Karl (2019) which was used with adults.

Our adaptation was presented on a PowerPoint presentation consisting of 28 slides, which we developed specifically to closely follow the administration guidelines from the original and revised face-to-face protocols (Gagarina et al., 2012, 2019). This task involves a scripted story which is presented alongside six pictures showing key events of the story. The narration involves four characters: a cat, a butterfly, a ball and a boy, and aims to assess macrostructural components of storytelling, such as structural complexity and use of internal state terms, as well as story comprehension. The activity involves three main parts: listening, retelling and comprehension. During the listening part of the activity, the participant is shown three colored envelopes and they are asked to pick one to reveal the story hidden inside. After opening one of the envelopes, the child looks at the pictures of the story which are initially presented on the screen silently. After examining the pictures, the child listens to the story, which is delivered in parallel with the relevant pictures. In our adaptation, participants all listened to the same pre-recorded version of the Cat story, which was presented in three sections, each accompanied by the relevant pictures. The story was recorded in advance by a female native English speaker. This was done in order to minimize variability amongst instances of task presentation in our participant sample. After listening to the story, participants are asked to retell the story. In our adaptation, the experimenter instructed participants by saying: "Now I want you to tell the story. Look at the pictures, and tell the best story you can. Imagine that you are telling the story to your favorite teacher who cannot see the pictures." These instructions were chosen to mimic the face-to-face protocol in which the experimenter doesn't share the story context with the participant (i.e., the experimenter cannot see the pictures). The adaptation intended to maintain this aspect of the task, as the availability of shared information can affect referential elements of children's retell (Gagarina et al., 2012, 2019). Consequently, we chose to address the retell toward a "favorite teacher," as we expected this to provide a friendly, but neutral figure, which also projects some elements of authority. For some of the adults with Down syndrome who were out of education this instruction was not applicable. In these instances, we asked them to "Imagine you are telling the story to a friend who cannot see the pictures." After hearing the instructions, the participant would begin telling their story as the experimenter moved through the pictures on the screen. As it was difficult at times to judge when the participant had completed the retelling of a portion of the story, the experimenter would check before moving to the next set of pictures by asking "Anything else?" The final part of the task assessed comprehension. In this portion, participants were shown the six pictures on one slide and asked questions about the story. For each question, the relevant pictures were highlighted by a red border, to focus the participants' attention to portions of the story relevant to the question. The activity was recorded for later transcription and scoring.

The adaptation used in the current study borrowed elements from the adaptations mentioned above, though it also included some additions. In particular, a few blank slides were added at intermediate points between the listening, retelling and comprehension portions of the task. This was done to ensure the participant was listening to the experimenter giving the instruction, rather than being distracted by the pictures on the screen. In particular, for the retelling portion, this addition ensured that the participant listened to the entire instruction before beginning to tell their story. A further addition included another blank slide placed at the beginning of the presentation which only contained an audio file playing a snippet of the recorded story (i.e., the speaker saying "1 day"). This was included to check that the sound from the presentation was audible to the participant through screen sharing before beginning the task. The slide was added after a few instances during task administration in which some participants were unable to hear the sound coming from the presentation once the first part of the recorded story was played out. The issue was most often due to the experimenter forgetting to tick the "Share sound" option when starting to screen share, while in a few cases it was due to an error of the PowerPoint presentation (e.g., presentation freezing). Slide numbers were also added to the presentation in order to facilitate administration for the experimenters.

Background measures

BPVS-3

Vocabulary comprehension was assessed using the British Picture Vocabulary Test, Third Edition (BPVS-3, Dunn and Dunn, 2009). The task was administered according to manual instructions and pictures from the stimulus book were shown

on screen using a visualiser camera. The video from the camera was screenshared with the participant by using the camera's visualiser application, as this allowed both the visualiser and the experimenter's camera to be active at the same time. During the teaching portion of the test (i.e., on Training plate A), participants' attention was drawn to the labels of the four pictures by saying: "Look, here are four pictures. There's picture One (pointing to the picture), picture Two (pointing to the picture), picture Three (pointing to the picture) and picture Four (pointing to the picture)." Whenever possible, participants were encouraged to verbalize their responses. The experimenter asked them to "Tell me which picture goes with the thing I have said." If participants tried to point or verbalized a label for the picture (e.g., "it's this one"), the experimenter would say "Could you tell me the number of the picture?" In some cases, the participants preferred to point rather than to verbalize their responses. In these instances, the experimenter asked the parent to assist and relay the number of the picture chosen by the participant.

KBIT-2 matrices

Non-verbal reasoning was assessed using the Matrices subtest from the Kaufman Brief Intelligence Test, Second Edition (KBIT-2, Kaufman and Kaufman, 2004). The assessment was administered as per manual instructions and pictures from the stimulus book were shown on screen by using a visualiser camera, in the same way as for the BPVS-3. The experimenter would administer the items by pointing to the relevant parts of the stimulus page as instructed by the manual. Again, participants were encouraged whenever possible to verbalize their responses by saying the letter associated with the chosen picture. If necessary, the experimenter would draw participants' attention to the labels of the response options by pointing out the letters associated with each picture during the teaching phase. If participants tried to point or verbalized a label for the picture (e.g., "it's the truck"), the experimenter would say "Could you tell me the letter of the picture?" In cases where participants preferred to point rather than verbalizing their responses, the experimenter asked the parent to relay the letter of the option chosen by the participant.

TROG-2

Grammar comprehension was assessed using the Test for Reception of Grammar, Second Edition (TROG-2; Bishop, 2003). The test was administered according to manual instructions and pictures from the stimulus book were shown on screen by using a visualiser camera, in the same way as for the BPVS3. In a similar fashion as for the BPVS-3, experimenters would point out the labels of the pictures during the teaching phase and encourage participants to verbalize their responses. When this was not possible, the parent would be asked to relay the number of the picture selected by the participant.

General procedure

All assessments were administered over a video call. Participants were invited to join a video call with the experimenter and were often accompanied by a parent, who sat next to them throughout the assessment and aided when required.

Data collection was typically completed over three video calls. The first video call arranged entailed completing the informant questionnaires with the parent or caregiver. The parental video call was typically scheduled first to create rapport with the parent and give them and the participant an opportunity to become more familiar with the researcher and ask questions about the study.

During the second and third video calls, participants completed the assessments administered by the experimenter. Participants with typical development completed the experimental measures during these sessions, while participants with Down syndrome were also administered the full battery of background measures (BPVS-3, TROG-2 and KBIT-2). The order of completion for both parental questionnaires and participant assessments was counterbalanced across participants: every other participant completed assessments in the reverse order, and the parent completed the questionnaires in the reverse order as well. When assessments were administered in the reversed order, BPVS-3 was moved to the beginning of the second session. This was done to maintain a standardized assessment at the beginning of each session, as this task provided a straightforward activity to engage the participant at the start of the call, before completing the more demanding experimental tasks.

Assessments were administered by the experimenters using a laptop or personal computer (PC) and a video camera (either USB or built-in webcam), using commercially available videoconferencing software (i.e., Microsoft Teams, Zoom). Experimenters wore headphones with a microphone during administration of the assessments. Administration of the standardized assessments was carried out using an OKIOLABS OKIOCAM T Compact A3 Visualiser/Document Camera.

Participants joined the testing sessions and completed the assessments using the equipment that was available to them in their home environment, typically either a laptop or tablet. In some cases, participants wore headphones, though the majority completed the assessments without them. This was especially the case for participants with Down syndrome: all but one of the participants preferred to not use headphones (these were sometimes not available, or the participant refused), though they always were encouraged to wear their hearing aids when available.

Scoring and data analysis

The speech samples collected during the MAIN retell activity were each transcribed by two separate independent transcribers.

Transcripts generated for each of the participants were compared to one another and conflicts resolved in accordance between the two original transcribers where possible, or by the first transcriber. Inter-transcriber reliability was checked for 50% of transcripts revealing an inter-rater agreement rate of 94–95%.

Coding and scoring of the narrative data were carried out by the first author. Scoring followed official guidelines from the MAIN assessment tool. The task is divided into three episodes, each involving two characters of the story. Each episode is structured to include five components: (1) an internal state generating the sequence of events (e.g., the cat saw the butterfly), followed by a (2) goal (e.g., the cat wanted to get the butterfly), followed by an (3) attempt (e.g., the cat jumped) and an (4) outcome (e.g., the cat fell in the bush). The sequence ends with an (5) internal state as a reaction to the outcome of the episode (e.g., the cat was hurt). Participants are awarded one point on story structure for each of the possible structural elements produced during their retell (up to 5 points for each episode). Two additional points are awarded for specifying the initial setting in terms of place and time (e.g., 1 day, at the lake). Following this scoring system, a score was generated for the number of structural elements present in the participants narratives. A score for use of internal state terms (ISTs) was also calculated by counting the total number of IST tokens produced by the participant during retell. Each use of words expressing perceptual states (e.g., look, see, hear), physiological states (e.g., hungry, tired, hurt), emotional states (e.g., sad, happy, scared), consciousness terms (e.g., awake, asleep), mental verbs (e.g., want, think, know), and linguistic verbs (e.g., say, shout, ask) was awarded 1 point. Finally, a comprehension score was obtained based on the participants' answers to a series of comprehension questions asked after retell, where each appropriate answer was awarded 1 point (for a maximum of 10 points). The comprehension questions are designed to tap into the participant's knowledge of characters' intentions and internal states.

As measures of narrative length and lexical diversity, respectively, the total number of words (TNW) produced by each participant during retell and the number of different words (NDW) used were also calculated. Total number of words included a count of all words produced by the participant during retell, with contractions (e.g., don't) being counted as two separate words. Only material relevant to the story was included in this count. Number of different words was calculated by counting each instance of different words occurring in retell, with same-stem words (e.g., run and ran, or want and wanted) being counted only in the first instance.

Statistical analyses were carried out in RStudio (RStudio Team, 2020, version 2022.02.3+492) using R (R Core Team, 2021, version 4.1.2). Parametric independent samples Student's t-tests were used to investigate mean differences between the groups on measures of story structure, comprehension, number of IST tokens, total words and number of different words. All assumptions of the Student's t-test were met. Besides t-statistics and p-values, we report effect sizes for each pair of comparisons as calculated using Cohen's d and interpreted according to the following commonly adopted guidelines (Cohen, 1988): small effect ($d = 0.2$); medium effect ($d = 0.5$); large effect ($d = 0.8$).

Two participants with Down syndrome (aged 27;2 and 30;4) were excluded from the analyses as they were not able to complete the MAIN retelling task, resulting in sample sizes of 11 participants with Down syndrome and 12 TD controls. An independent samples Student's t-test confirmed that there was no significant difference between the mean chronological age of TD participants and the mean vocabulary age equivalent for the participants with Down syndrome [$t(23) = -0.08$, $p = 0.937$].

Results

Supplementary Table 2 reports means (standard deviations) and ranges for Macrostructural (Story Structure and Internal State Terms) and Comprehension measures derived from the MAIN retell activity scoring, as well as means (and standard deviations) for narrative length (TNW) and lexis (NDW) produced by the two groups.

Statistical analyses revealed no significant difference between the groups on narrative length, as measured by total number of words produced during retell [$t(21) = -0.877$, $p = 0.39$, $d = 0.366$], however a significant difference in lexis emerged when comparing the number of different words produced by the groups [$t(21) = -2.30$, $p < 0.05$, $d = -0.961$]. Typically developing children produced a greater quantity of different words throughout their narratives relative to the participants with Down syndrome, despite overall length showing no difference across the groups.

Comparisons of macrostructural performance revealed a significant difference between the groups on the Story Structure score [$t(21) = -2.68$, $p < 0.05$, $d = -1.12$], with TD controls achieving higher scores compared to the participants with Down syndrome This indicates that TD children tended to include more elements of story grammar in their retells, such as initiating events, reactions, characters' intentions, actions and consequences. No difference, however, was found in the number of internal state term (IST) tokens produced by the two groups [$t(21) = -1.55$, $p = 0.135$, $d = -0.649$], suggesting similar rates of acknowledgment of characters' emotional, physiological or mental states across the groups.

Finally, a significant difference emerged when comparing Comprehension scores [$t(21) = -3.03$, $p < 0.01$, $d = -1.26$], indicating that participants with Down syndrome were outperformed by the TD group on our measure of story comprehension. This indicates lower overall accuracy of responses to questions in our study for the group with DS relative to our TD controls.

Discussion

The current paper presents the methodological approach and findings of a pilot study of narrative abilities in adolescents and young adults with Down syndrome. The study employed a picture-based retelling activity from the MAIN narrative task developed by Gagarina et al. (2019), as well as a range of background measures of language and cognitive ability, all presented online due to the COVID-19 epidemiological restrictions in place at the time of recruitment. The analyses included in this report sought to investigate aspects of macrostructural abilities, comprehension, narrative length and lexical diversity of story retell samples produced by a group of participants with Down syndrome aged between 15 and 33 years and a group of vocabulary age-matched TD controls. Due to the small sample size, the results should be interpreted with caution. However, considering the dearth of studies using online methodology to investigate language skills in this vulnerable population, we believe that our insights will provide valuable guidance for future research, particularly in relation to the assessment of language skills in the population with Down syndrome.

Our analyses revealed differences between the participants with Down syndrome and the vocabulary-matched controls in terms of both narrative structure and narrative comprehension. Thus, in answer to our second research question, our results clearly indicate that the adolescents and adults with Down syndrome, aged between 15;3 and 32;11, were outperformed by their much younger TD counterparts, aged 4;4 to 10;6. With regards to story structure, our control participants were able to produce more story components, i.e., Setting, Goals, Attempts, Outcomes, initiating events and reactions, in their retells compared to the adolescents and young adults with Down syndrome. This suggests that individuals with Down syndrome exhibit more difficulty with structural aspects of story-telling when vocabulary ability is controlled for. Such a difference may be attributed to overall expressive language deficits in the population with Down syndrome, widely reported in the literature. However, in addition to the recognized syntactic deficits (also confirmed by our participants' poor performance on the standardized measure of grammar comprehension, discussed in more detail below), it is possible that our participants were less skilled in producing relevant story components due to an additional presence of pragmatic deficits. Taking into account the perspective of the audience and the common ground shared between the speaker and listener is crucial in telling a successful narrative. Pragmatic difficulties have been observed in children with Down syndrome (Smith et al., 2017), despite early reports of relatively spared pragmatics (Roberts et al., 2007). A more detailed investigation of our participants' pragmatic abilities, independently of the narrative, may shed light on the nature of this finding.

The same issue is pertinent with regards to our second finding, our participants' poorer performance on story comprehension. Adolescents and young adults with Down syndrome provided fewer accurate answers than TD matches when asked a series of comprehension questions tapping into character's internal states and goals. The finding that participants with Down syndrome performed below vocabulary age expectations on the comprehension portion of the task fits with the widely reported weaknesses in sentence comprehension in this population. While the current paper does not focus on microstructural components of the narratives produced, our participants' poor comprehension score is in line with the literature highlighting sentence comprehension as a particular weakness in Down syndrome. As observed in their poor overall scores on the standardized measure of grammar comprehension administered in the current study, TROG-2, our participants showed significant difficulties interpreting a range of syntactic structures incorporated in this assessment. This is in line with existing literature showing floor scores on TROG-2 across different ages of individuals with DS (e.g., Frizelle et al., 2019), as well as studies focusing on specific complex syntactic structures that include passives (Ring and Clahsen, 2005; Perovic and Wexler, 2019) or relative clauses (Joffe and Varlokosta, 2007; Frizelle et al., 2019). However, it is not clear how much of the comprehension difficulty shown by our participants can be attributed to their grammatical difficulties, compared to possible difficulties in understanding mental states of the characters involved in the story, as intimated earlier.

Our analysis did not reveal differences between the groups on the number of internal state term (IST) tokens—words used to express the mental and emotional states of characters in the story—present in each group's narratives. This finding indicates that participants with Down syndrome produced words expressing internal states at a level comparable to TD controls matched on receptive vocabulary, suggesting that possible syntactic and/or ToM difficulties associated with the expression of mental states did not affect their production beyond vocabulary-based expectations. However, a larger sample of participants is needed to help us establish these facts: despite the absence of statistically significant difference between the groups, participants with Down syndrome showed lower performance on this measure. Nonetheless, a more granular analysis of the types of IST tokens and associated syntactic constructions produced by speakers in our groups may provide insights into their relative abilities to interpret and express the goals and emotions of others.

Interestingly, we found no evidence of a difference in narrative length between the two groups, as calculated by the total number of words produced by participants. This suggests that the difficulty exhibited by participants with Down syndrome in the production of structural components cannot be explained in terms of the raw length of their narratives alone. In other words, despite producing similar amounts of

words, participants with DS still mentioned fewer elements of the story structure during retell. While narrative length was similar between the groups, the number of different words used differed significantly: individuals with Down syndrome were outperformed in this respect, indicating that TD participants exhibited greater levels of lexical diversity compared to participants with Down syndrome. This may suggest that participants with DS tended to focus on selected aspects of the story to the exclusion of others, possibly reformulating or repeating information more often throughout their narratives.

Our results are not in line with those of studies that included younger participants with Down syndrome reviewed earlier. Recall that children and adolescents with Down syndrome (10−16 years, mean age: 12) in Zanchi et al. (2021) produced story structure and event-based information at a level commensurate to their expected non-verbal ability and MLU. Previously, Miles and Chapman (2002) have suggested that adolescents and young adults with Down syndrome (12–26 years, mean age: 18.76) produce narratives with structural elements, episodic events and thematic content that are in line with syntax comprehension and non-verbal cognitive levels, but which surpass what might be expected on the basis of MLU. A possibility, such as the one advanced by Channell et al. (2015), is that conceptual story telling abilities are relatively spared in Down syndrome, and at the level of non-verbal ability, but expression is limited by underdeveloped syntactic abilities. This is supported by Finestack et al. (2012) who also report an advantage on macrostructural story components when controlling for non-verbal mental age, but not when controlling for MLU. As noted earlier however, MLU may not be an appropriate measure for capturing expressive syntax abilities in adults with Down syndrome, as the informativeness of MLU as a measure of syntactic complexity significantly declines with age, and after an MLU of 4.0 (CA: 3 years) has been reached (Scarborough et al., 1991). Martzoukou et al. (2020) had speculated on the effect of world knowledge, associated with age, on the structural organization of narratives in Down syndrome. They hypothesized that the greater world knowledge of adults with Down syndrome would translate into better narrative structure compared to that of younger TD controls matched for expressive vocabulary and non-verbal mental age. However, this hypothesis did not find support in their analysis and does not seem to be supported by the analyses presented in our study, at least in terms of receptive vocabulary. Overall, our findings are partly in line with results reported in Martzoukou et al. (2020), the only other study employing the same instrument to assess narrative skills of adults with Down syndrome, though Greek rather than English-speaking, and administered in person rather than online. The results of their study also revealed poorer performance on both story structure and story comprehension from adults with Down syndrome (19–46 years, mean age 28;2) when compared to expressive vocabulary-matched TD controls. In terms of use of IST tokens, our study found no differences in the frequencies of internal state terms use between the two groups, while Martzoukou et al. report significantly fewer ISTs in the productions of their participants with Down syndrome compared to both of their TD control groups, one matched on expressive vocabulary and the other on non-verbal ability. Martzoukou et al. interpret the lower frequencies of mental state terms in the narratives of participants with Down syndrome as a result of their poor syntactic ability: since verbs used to express internal state terms almost always require complex syntactic constructions such as complement clauses (de Villiers and de Villiers, 2009), it may be this syntactic complexity that precludes use of mental state terms in individuals with Down syndrome. However, this difference in IST use might alternatively be explained by the relatively older age of the sample recruited by Martzoukou et al. It is possible that, given the ages reported in their study, some of the participants included in their sample might have already started to experience symptoms of cognitive decline which reflected on their language performance. Crucially in our sample, we include participants below the age of 35, a stage of life at which we expect language development to be approaching adult form. We would expect individuals with Down syndrome between the ages of 15 and 35 to show degrees of variability in their relative skills, across both language and cognition, as performance in this population is often reported to be highly heterogeneous (Roberts et al., 2007). However, we wouldn't expect neuropsychological symptoms of Alzheimer's disease to influence performance in this age range, as the impact of dementia-related decline becomes manifest around the fourth decade of life (Ballard et al., 2016). Based on this, we would expect the abilities of our sample to be reflective of adult macrostructural narrative abilities, unaffected by decline.

With regards to our third research question, in this study we were able to show that diverse types of data can be successfully obtained *via* online remote administration from individuals with Down syndrome. Our results suggest that measures of general language and cognitive abilities can be used remotely with the population with DS. Only two participants with Down syndrome (aged 27;2 and 30;4) were excluded from the analyses reported as they were not able to complete the MAIN retelling task, however, they completed the background measures. These participants' inability to complete the retell task was primarily due to limited expressive language abilities and use of prompting from the caregiver during task administration.

All participants completed the grammar and vocabulary comprehension tasks, in addition to the task assessing non-verbal reasoning. Here we discuss how our participants scored on these very same measures as administered online, compared to those reported in previous literature, but administered face-to-face. While direct comparisons cannot be made, due to different age ranges of participants involved, the mean scores on BPVS-3 and TROG-2 for our sample of participants seem in line with those reported for in person assessments in the literature, suggesting that that online assessment of vocabulary

and grammar may be viable in the population with Down syndrome. Our participants' scores on TROG-2 are similar to those reported in the previously reviewed Finestack et al. (2012) for adolescents and young adults with Down syndrome aged between 12 and 26 mean age 18.76. The mean TROG-2 raw score reported in their study, 2.63 (SD = 1.58), range 0–6, is comparable to that seen in our participants, mean 3.38, SD (2.53), range: 0–7. With regards to BPVS-3, adults with DS from a LonDownS Consortium study (Startin et al., 2019b), aged between 19 and 59 (mean age 36.47 years), obtained a mean raw score of 95.94 (SD: 31.99) (range: 38–158), which is comparable to our participants' mean score of 89.23 (26.03) range: 53–119. Similar levels of non-verbal ability have also been previously observed on KBIT-2: in another LonDownS Consortium study, Startin et al. (2016) report mean non-verbal raw score of 14.98 (6.90), range 0–32, for a large sample of young healthy adults with DS aged 16–35 years (mean age 25.24 years), which is in line with our sample's mean score of 15.31 (5.06), range: 2–21. In addition, the finding that narrative length was comparable across our participant samples when matched on age of vocabulary comprehension suggests that the online elicitation approach was similarly effective in eliciting narrative discourse from both participants with DS and TD controls remotely. Further analyses of the initial data presented in the current paper will explore the nature of this finding in more details, as our results suggest that despite producing a comparable number of words, participants with DS included fewer elements of story structure and used a more limited range of different words in their narratives. We provide a more detailed discussion of our general experience with the remote collection of data in the section below, to allow future researchers to make informed choices when considering remote methods of assessment.

Methodological considerations on remote administration of experimental materials

Online methods of data collection in the population with Down syndrome present both advantages and disadvantages which must be taken into consideration when designing remote research approaches. Online methods minimize the need for travel and significantly reduce contingencies associated with the costs and time-investments of travel, for both researchers and participants. Furthermore, they allow for a relatively comfortable mode of testing, thanks to the commercial availability of numerous videoconferencing tools which have seen a significant uptake in usage in recent times, especially due to the social-isolation restrictions put in place all over the world. Participants can then be tested at home, from a familiar environment, while reducing possible health risks associated with face-to-face contact.

In terms of participant engagement, the pilot has yielded promising results so far. Our experiences do not confirm worries relating to participants' ability to engage in online activities. The families involved in the study were always able to connect to the videocalls using their electronic devices. In some cases, it was the child or adult participant who provided help to the parents, who were sometimes less familiar with online videoconferencing tools. In other cases, participants did require help from a parent to join the calls (this was virtually always the case for the TD children, but also for some of our participants with Down syndrome), but they were nonetheless able to engage with the experimenter and participate in the online tasks. For two participants with Down syndrome, the narrative task presented particular challenges, likely associated with more pronounced impairments in expressive skills. We speculate that the task itself may have been too demanding for these participants, however, it cannot be excluded that a face-to-face setting may have facilitated their performance. Future research would benefit the field by examining performance differences in people with intellectual disability between face-to-face and remote online task administration. As for recruitment, we again cannot exclude that the online nature of the study might have dissuaded some families from participating, however, we can report that a number of families expressed relief at the notion of being able to participate from home. Such relief might be in part associated with the relative ease of remote participation on the side of families and seems reasonable given the broader global context in which data collection has been taking place. In this respect, remote online assessment represents a promising new approach to data collection, especially when working with extremely vulnerable populations such as people with Down syndrome. Furthermore, an added benefit of the remote approach to assessment was that of allowing us to reach and involve participants and families based all over the United Kingdom, while significantly reducing traveling costs incurred by researchers.

Alongside the promising outcomes for both recruitment and data collection, we must also consider some of the drawbacks of remote online assessment. One of the main issues involve the representativeness of our sample, due to possible disparities in accessing the necessary technology to take part in online studies. While the global pandemic forced schoolchildren and students all over the world to participate in online schooling during official lockdowns in 2020 and 2021, forcing them and their caregivers to invest into technology, this was not possible for many families who faced economic hardship (UNICEF, 2021).

With regard to the experimental set up itself, the central issues revolve around a loss of experimental control over the surrounding environment, the equipment used, and a lack of direct influence over the engagement of the participant. In our pilot study, we observed significant variation in terms of the equipment available to the participants, as they completed the sessions from their home environment. Participants typically

completed assessments using a laptop or tablet, in most cases without wearing headphones, as these were uncomfortable to them or not available to the families. The majority of participants, across both groups, completed the assessments with a parent or caregiver present who provided assistance when necessary, without interfering with the assessments. Variations in the type of equipment available to participants is an important consideration in the design of online elicitation studies, as it may play a role in how the participants are able to engage with the tasks: for example, features such as screen size and resolution of the electronic device used could affect the visibility of the materials presented.

Most participants were able to complete the assessments following instructions given by the experimenter and did not require further assistance from the caregiver. In some cases, the caregiver helped by redirecting the participant's attention toward the tasks when distracted by environmental stimuli (e.g., noise, other family members in the house). In other cases, specifically with two participants with Down syndrome, the help of the parent was actively required in administering the background assessments, as the participants chose to point to pictures rather than verbalizing their responses. Given that the assistance, or at the very least the presence, of a caregiver is likely necessary during task administration, we recommend that future studies take measures to instruct the parent on how to behave and assist during the assessments, in order to avoid interference that can invalidate the quality of the data collected.

Another potential issue of remote online assessments, especially relevant when eliciting language samples, relates to the audio quality and intelligibility of the recordings collected. In this study, we collected language samples from the MAIN retell activity (Gagarina et al., 2019), which were subsequently transcribed and scored. For most participants, the reduced quality of language samples collected without using headphones did not have a significant impact on the experimenters' ability to transcribe the samples. However, this was not always the case for participants with Down syndrome, for whom in a few cases, low audio quality coupled with intelligibility issues significantly affected the ease of transcription. Our inter-transcriber reliability rates were excellent, however, providing families with equipment (particularly headphones and a microphone) could contribute to minimizing audio quality issues and may improve the intelligibility of the speech samples collected. One important drawback of using headphones, however, is that the parent or caregiver assisting with the session would not have auditory access to the instructions given by the experimenter. A further aspect that merits consideration relates to hearing difficulties, which appear to be common in the population with DS (Shott, 2000). In our study, six participants reported some degree of hearing loss which could affect their task performance. Of these, two were excluded from analysis as they failed to complete the retelling activity, while the remaining four wore hearing aids during task administration. While the use of hearing aids may help minimize the impact of hearing difficulties, remote language elicitation designs should seek to account for this factor: though this wasn't included in the present design, language research would benefit from the adaption of a hearing screening procedure for remote use.

Finally, we experienced minimal issues with regards to connection quality during videocalls, another area of uncertainty when administering assessments online. Participants overall seemed to have access to good enough internet connection to allow completion of the activities in the absence of signal degradation that could significantly affect task administration. Occasional issues were observed in terms of audio quality, with a few instances of audio glitches. Because the assessments required the participant being able to hear speech produced by the experimenter, we allowed experimenters to repeat items if the participant indicated that they did not hear the experimenter's prompt, though this was seldom the case for the tasks presented above (in particular, this was especially relevant for language comprehension tasks such as BPVS-3 and TROG-2).

Conclusion

The current research contributes to the growing body of literature documenting the language skills of adults with Down syndrome, a relatively understudied demographic, by focusing on macrostructural narrative language elicited by means of a novel methodological approach. In the first entirely remote study of narrative language abilities in adults with Down syndrome, we adapted the assessment tool for remote online use over videoconference. In addition to contributing valuable insights into the feasibility of remote online research designs with participants with intellectual disability, the study is also the first to assess narrative language in a group of adults with Down syndrome at an age range where language abilities are approaching adult performance, while unlikely to be affected by cognitive deterioration associated with Alzheimer's disease. We report a disadvantage of participants with Down syndrome relative to TD controls matched on age of vocabulary comprehension on global measures of story structure and story comprehension, as well as lexical diversity, though the groups did not differ on story length as measured by total number of words. In the current report, we discuss the implications of such findings in relation to previous literature assessing macrostructural narrative skills of children and adolescents with Down syndrome, as well as the more sparse evidence available on adults, and reflect on the outcomes of our remote online approach to language assessment. We conclude that remote online methodological approaches are viable tools of eliciting speech samples and assessing expressive language skills in adolescent and young adults who have Down syndrome.

Data availability statement

The datasets presented in this article are not readily available because this dataset includes data from vulnerable participants. Requests to access the datasets should be directed to Elisa Mattiauda, elisa.mattiauda.18@ucl.ac.uk.

Ethics statement

The studies involving human participants were reviewed and approved by UCL Research Ethics. Written informed consent to participate in this study was provided by the participants and their legal guardians. Child participants provided assent through a child-friendly form.

Author contributions

EM, AH, and AP designed the study. EM collected and analyzed the data and wrote the manuscript. AP contributed to the writing of the manuscript. AH edited the manuscript. All authors approved the final version.

Funding

The Baily Thomas Charitable Fund, Doctoral Fellowship Award number: 5683-8804, to EM.

Acknowledgments

We would like to thank our participants and their families, and charities that helped with the recruitment process. We also thank the Baily Thomas Charitable Fund for providing the funds to conduct this research.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.841543/full#supplementary-material

References

Abbeduto, L., Murphy, M. M., Cawthon, S. W., Richmond, E. K., Weissman, M. D., Karadottir, S., et al. (2003). Receptive language skills of adolescents and young adults with Down or fragile X syndrome. *Am. J. Mental Retardation* 108, 149–160. doi: 10.1352/0895-8017(2003)108<0149:RLSOAA>2.0.CO;2

Ash, S., Moore, P., Vesely, L., and Grossman, M. (2007). The decline of narrative discourse in Alzheimer's disease. *Brain Lang.* 103, 8–249. doi: 10.1016/j.bandl.2007.07.105

Ballard, C., Mobley, W., Hardy, J., Williams, G., and Corbett, A. (2016). Dementia in Down's syndrome. *Lancet Neurol.* 15, 622–636. doi: 10.1016/S1474-4422(16)00063-6

Berman, R., and Slobin, D. (1994). *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Hillsdale, NJ: Erlbaum.

Bishop, D. (2003). *The Test for Reception of Grammar - Version 2*. London: Psychological Corporation.

Bruner, J. (1986). *Actual Minds, Possible Worlds*. Cambridge, MA: Harvard University Press. doi: 10.4159/9780674029019

Butler, L., La Valle, C., Schwartz, S., and H. Tager-Flusberg. (2022). Remote natural language sampling of parents and children with autism spectrum disorder: Role of activity and language level. *Front. Commun.* 7, 820564. doi: 10.3389/fcomm.2022.820564

Campbell, C., Landry, O., Russo, N., Flores, H., Jacques, S., and Burack, J. A. (2013). Cognitive flexibility among individuals with Down syndrome: Assessing the influence of verbal and nonverbal abilities. *Am. J. Intellect. Dev. Disabil.* 118, 193–200 doi: 10.1352/1944-7558-118.3.193

Carter Young, E., and Kramer, B. M. (1991). Characteristics of age-related language decline in adults with Down syndrome. *Ment. Retard.* 29, 75–79.

Caselli, M. C., Monaco, L., Trasciani, M., and Vicari, S. (2008). Language in Italian children with Down syndrome and with specific language impairment. *Neuropsychology* 22:27. doi: 10.1037/0894-4105.22.1.27

Channell, M. M., McDuffie, A. S., Bullard, L. M., and Abbeduto, L. (2015). Narrative language competence in children and adolescents with Down syndrome. *Front. Behav. Neurosci.* 9:283. doi: 10.3389/fnbeh.2015.00283

Chapman, R. S., and Hesketh, L. J. (2000). The behavioral phenotype of individuals with Down syndrome. *Mental Retardation Dev. Disabil. Res. Rev.* 6, 84–95. doi: 10.1002/1098-2779(2000)6:2<84::AID-MRDD2>3.0.CO;2-P

Chapman, S. B., Highley, A. P., and Thompson, J. L. (1998). Discourse in fluent aphasia and Alzheimer's disease: Linguistic and pragmatic considerations. *J. Neurolinguistics* 11, 55–78. doi: 10.1016/S0911-6044(98)00005-0

Cleave, P., Bird, E. K. R., Czutrin, R., and Smith, L. (2012). A longitudinal study of narrative development in children and adolescents with down syndrome. *Intellect. Develop. Disabil.* 50, 332–342. doi: 10.1352/1934-9556-50.4.332

Clift, A. K., Coupland, C., Keogh, R. H., Hemingway, H., and and, J., Hippisley-Cox (2021). COVID-19 mortality risk in Down syndrome: Results from a cohort study of 8 million adults. *Ann. Intern. Med.* 174, 572–576. doi: 10.7326/M20-4986

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

Cooper, S. A., and Collacott, R. A. (1995). The effect of age on language in people with Down's syndrome. *J. Intellectual Disabil. Res.* 39, 197–200. doi: 10.1111/j.1365-2788.1995.tb00501.x

de Graaf, G., Buckley, F., and Skotko, B. G. (2021). Estimation of the number of people with Down syndrome in Europe. *Eur. J. Hum. Genet.* 29, 402–410. doi: 10.1038/s41431-020-00748-y

de Villiers, J.G., and de Villiers, P. A. (2009). "Complements enable representation of the contents of false beliefs: The evolution of a theory of theory of mind," in *Language Acquisition. Palgrave Advances in Linguistics*, ed S. Foster-Cohen (London: Palgrave Macmillan).

de Villiers, J. G., and de Villiers, P. A. (2000). "Linguistic determinism and false belief," in: *Children's Reasoning and the Mind*, eds P. Mitchell and K. Riggs (Hove: Psychology Press).

Devenny, D. A., and Krinsky-McHale, S. (1998). Age-associated differences in cognitive abilities in adults with Down syndrome. *Top. Geriatr. Rehabil.* 13, 65–72. doi: 10.1097/00013614-199803000-00008

Devenny, D. A., Krinsky-McHale, S. J., Sersen, G., and Silverman, W. P. (2000). Sequence of cognitive decline in dementia in adults with Down's syndrome. *J. Intellectual Disabil. Res.* 44, 654–665. doi: 10.1111/j.1365-2788.2000.00305.x

Dunn, L., and Dunn, D. (2009). *The British Picture Vocabulary Scale, 3rd Ed.* Brentford: GL Assessment.

Eadie, P. A., Fey, M. E., Douglas, J. M., and Parsons, C. L. (2002). Profiles of grammatical morphology and sentence imitation in children with specific language impairment and Down syndrome. *J. Speech Lang. Hear. Res.* 45, 720–732. doi: 10.1044/1092-4388(2002/058)

El-Raziq, M. (2022). *On-line LITMUS-MAIN assessment in Arabic-speaking children with and without ASD: Challenges and opportunities*. Paper presented at the ZAS Meeting "Online elicitation of narrative texts: Summarizing experience and making plans".

Finestack, L. H., Palmer, M., and Abbeduto, L. (2012). Macrostructural narrative language of adolescents and young adults with Down syndrome or fragile X syndrome. *Am. J. Speech Lang. Pathol.* 21, 29–46. doi: 10.1044/1058-0360(2011/10-0095)

Fowler, A., Gelman, R., and Gleitman, L. (1994). "The course of language learning in children with Down syndrome," in: *Constraints on Language Acquisition: Studies of Atypical Children*, ed H. Tager-Flusberg (Hillsdale, NJ: L. Erlbaum Associates).

Frizelle, P., Thompson, P. A., Duta, M., and Bishop, D. (2019). The understanding of complex syntax in children with Down syndrome. *Wellcome Open Res.* 3:140. doi: 10.12688/wellcomeopenres.14861.2

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Bohnacker, U., et al. (2019). MAIN: multilingual assessment instrument for narratives – revised. *ZAS Papers Linguistics* 63:e516. doi: 10.21248/zaspil.63.2019.516

Gagarina, N. V., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiuniene, I., et al. (2012). MAIN: Multilingual assessment instrument for narratives. *ZAS Papers Linguistics* 56, 155–155. doi: 10.21248/zaspil.56.2019.414

Hesketh, L. J., and Chapman, R. S. (1998). Verb use by individuals with Down syndrome. *Am. J. Ment. Retard.* 103, 288–304.

Jarrold, C., and Baddeley, A. D. (2001). Short-term memory in Down syndrome: applying the working memory model. *Down Syndr. Res. Pract.* 7, 17–23. doi: 10.3104/reviews.110

Joffe, V., and Varlokosta, S. (2007). Patterns of syntactic development in children with Williams syndrome and Down's syndrome: evidence from passives and wh–questions. *Clin. Linguist. Phonetics* 21, 705–727. doi: 10.1080/02699200701541375

Justice, L. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L., and Gillam, R. B. (2006). The index of narrative microstructure: a clinical tool for analyzing school-age children's narrative performances. *Am. J. Speech-Lang. Pathol.* 15, 177–191. doi: 10.1044/1058-0360(2006/017)

Kapalková, S., Slančová, D., and Nemcová, M. (2021). "Multilingual assessment instrument for narratives protocols for online elicitation (Slovak version)," in: *Based on Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Bohnacker, U. & Walters, J. (2019). MAIN: Multilingual Assessment Instrument for Narratives. Revised version. ZAS Papers in Linguistics, 63. Slovak version. Translated*, eds S. Kapalková, D. Slančová, and M. Nemcová (Slovakia: Comenius University in Bratislava).

Karl, K. B. (2019). *Language Across Lifespans: Researching Linguistic Skills With a Focus on Narratives* [Presentation]. MAIN Text & Tea Meeting (2nd Edition), Online event (see: https://sites.google.com/view/text-tea/meetings). Protocol developed by Karl, K. B. (Ruhr-Universität Bochum, Germany).

Kaufman, A. S., and Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test–Second Edition*. Circle Pines, MN: AGS Publishing.

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, 1634. doi: 10.3389/fpsyg.2020.01634

Lai, F., and Williams, R. S. (1989). A prospective study of alzheimer disease in down syndrome. *Arch. Neurol.* 46, 849–853. doi: 10.1001/archneur.1989.00520440031017

Lautarescu, B. A., Holland, A. J., and Zaman, S. H. (2017). The early presentation of dementia in people with down syndrome: a systematic review of longitudinal studies. *Neuropsychol. Rev.* 27:e9341. doi: 10.1007/s11065-017-9341-9

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.* 62, 3982–3990. doi: 10.1044/2020_JSLHR-20-00202

Martin, G. E., Klusek, J., Estigarribia, B., and Roberts, J. E. (2009). Language characteristics of individuals with Down syndrome. *Topics Lang. Disord.* 29, 112–132. doi: 10.1097/TLD.0b013e3181a71fe1

Martzoukou, M., Nousia, A., and Marinis, T. (2020). Narrative abilities of adults' with down syndrome as a window to their morphosyntactic, socio-cognitive, and prosodic abilities. *Front. Psychol.* 11:e02060. doi: 10.3389/fpsyg.2020.02060

Matthews, D., Biney, H., and Abbot-Smith, K. (2018). Individual differences in children's pragmatic ability: A review of associations with formal language, social cognition, and executive functions. *Lang. Learn. Dev.* 14, 186–223. doi: 10.1080/15475441.2018.1455584

McCarron, M., McCallion, P., Reilly, E., Dunne, P., Carroll, R., and Mulryan, N. (2017). A prospective 20-year longitudinal follow-up of dementia in persons with Down syndrome. *J. Intellect. Disabil. Res.* 61, 843–852. doi: 10.1111/jir.12390

McCarron, M., McCallion, P., Reilly, E. S., and Mulryan, N. (2014). A prospective 14-year longitudinal follow-up of dementia in persons with Down syndrome. *J. Intellectual Disabil. Res.* 58, 61–70. doi: 10.1111/jir.12074

Miles, C. R., and Sindberg, H. (2006). Sampling context affects MLU in the language of adolescents with down syndrome. *J. Speech Lang. Hear. Res.* 49, 325–337. doi: 10.1044/1092-4388(2006/026)

Miles, S., and Chapman, R. S. (2002). Narrative content as described by individuals with down syndrome and typically developing children. *J. Speech Lang. Hear. Res.* 45, 175–189. doi: 10.1044/1092-4388(2002/013)

Norbury, C. F., and Bishop, D. V. M. (2003). Narrative skills in children with communication impairments. *Int. J. Lang. Commun. Impairments* 38, 287–313. doi: 10.1080/1368203100001081 33

Pearson, B., and De Villiers, P. (2006). *Discourse, Narrative and Pragmatic Development. Concise Encyclopedia of Pragmatics.* Oxford: Elsevier. doi: 10.1016/B0-08-044854-2/00841-5

Perovic, A. (2006). Syntactic deficit in Down syndrome: More evidence for the modular organisation of language. *Lingua* 116, 1616–1630. doi: 10.1016/j.lingua.2005.05.011

Perovic, A., and Wexler, K. (2019). "The effect of age on language in adults with intellectual disabilities: A comparison of passives in Down syndrome and Williams syndrome," in: *Interdisciplinary Linguistic and Psychiatric Research on Language Disorders. Faculty of Philosophy University of Zagreb and Psychiatry Clinic Vrapče*, eds V. Erdeljac and M. Sekulic Sovic (FF-Press). doi: 10.17234/9789531758314.03

Price, J. R., Roberts, J. E., Hennon, E. A., Berni, M. C., Anderson, K. L., and Sideris, J. (2008). Syntactic complexity during conversation of boys with fragile X syndrome and Down syndrome. *J. Speech Lang. Hear. Res.* 51, 3–15 doi: 10.1044/1092-4388(2008/001)

Pulsifer, M. B., Evans, C. L., Hom, C., Krinsky-McHale, S. J., Silverman, W., Lai, F., et al. (2020). Language skills as a predictor of cognitive decline in adults with Down syndrome. *Alzheimer's Dementia* 12:e12080. doi: 10.1002/dad2.12080

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ring, M., and Clahsen, H. (2005). Morphosyntax in Down's syndrome: is the extended optional infinitive hypothesis an option? *Stem Spraak Taalpathologie* 13, 3–13. Available online at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.956.2065&rep=rep1&type=pdf

Roberts, J. E., Price, J., and Malkin, C. (2007). Language and communication development in Down syndrome. *Ment. Retard. Dev. Disabil. Res. Rev.* 13, 26–35. doi: 10.1002/mrdd.20136

RStudio Team (2020). *RStudio: Integrated Development for R.* RStudio, PBC, Boston, MA.

Scarborough, H., Rescorla, L., Tager-Flusberg, H., Fowler, A., and Sudhalter, V. (1991). The relation of utterance length to grammatical complexity in normal and language-disordered groups. *Appl. Psycholinguist.* 12, 23–46. doi: 10.1017/S014271640000936X

Sealey, L. R., and Gilmore, S. E. (2008). Effects of sampling context on the finite verb production of children with and without delayed language development. *J. Commun. Disord.* 41, 223–258. doi: 10.1016/j.jcomdis.2007.10.002

Segal, A., and Pesco, D. (2015). Narrative skills of youth with Down syndrome: a comprehensive literature review. *J. Dev. Phys. Disabil.* 27, 721–743. doi: 10.1007/s10882-015-9441-5

Shott, S. R. (2000). Down syndrome: Common paediatric ear, nose and throat problems. *Quarterly* 5, 1–6.

Sinai, A., Bohnen, I., and Strydom, A. (2012). Older adults with intellectual disability. *Curr. Opin. Psychiatry* 25, 359–364. doi: 10.1097/YCO.0b013e328355ab26

Smith, E., N?ss, K. A. B., and Jarrold, C. (2017). Assessing pragmatic communication in children with down syndrome. *J. Commun. Disord.* 68, 10–23. doi: 10.1016/j.jcomdis.2017.06.003

Startin, C. M., Hamburg, S., Hithersay, R., Al-Janabi, T., Mok, K. Y., Hardy, J., et al. (2019a). Cognitive markers of preclinical and prodromal Alzheimer's disease in Down syndrome. *Alzheimer's Dementia* 15, 245–257. doi: 10.1016/j.jalz.2018.08.009

Startin, C. M., Hamburg, S., Hithersay, R., Davies, A., Rodger, E., Aggarwal, N., et al. (2016). The LonDownS adult cognitive assessment to study cognitive abilities and decline in Down syndrome. *Wellcome Open Res.* 1:e9961. doi: 10.12688/wellcomeopenres.9961.1

Startin, C. M., Hamburg, S., Strydom, A., and LonDown, S. (2019b). Comparison of receptive verbal abilities assessed using the KBIT-2 and BPVS3 in adults with Down syndrome. *Front. Psychol.* 9:2730. doi: 10.3389/fpsyg.2018.02730

Strauss, D., and Eyman, R. K. (1996). Mortality of people with mental retardation in California, 1986–1991. *Am. J. Mental Retardation* 100, 643–653.

Sultana, N. (2022). "Evaluating the potential and limits of online language assessment: Assessing the same children's narrative abilities in an online mode versus face-to-face mode," in: *Paper Presented at the ZAS Meeting Online Elicitation of Narrative Texts: Summarizing Experience and Making Plans.*

Ukrainetz, T. A., Justice, L. M., Kaderavek, J. N., Eisenberg, S. L., Gillam, R. B., and Harm, H. M. (2005). The development of expressive elaboration in fictional narratives. *J. Speech Lang. Hear. Res.* 48, 1363–1377. doi: 10.1044/1092-4388(2005/095)

UNICEF (2021). *Digital Learning for Every Child*. UNICEF: United Nations Children's Fund.

Villani, E. R., Vetrano, D. L., Damiano, C., Paola, A. D., Ulgiati, A. M., Martin, L., et al. (2020). Impact of COVID-19-related lockdown on psychosocial, cognitive, and functional well-being in adults with down syndrome. *Front. Psychiatry* 11:1150. doi: 10.3389/fpsyt.2020.578686

Wu, J., and Morris, J. K. (2013). The population prevalence of Down's syndrome in England and Wales in 2011. *Eur. J. Human Genet.* 21, 1016–1019. doi: 10.1038/ejhg.2012.294

Zanchi, P., Zampini, L., and Panzeri, F. (2021). Narrative and prosodic skills in children and adolescents with Down syndrome and typically developing children. *Int. J. Speech Lang. Pathol.* 23, 286–294. doi: 10.1080/17549507.2020.1804618

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 14 October 2022
DOI 10.3389/fpsyg.2022.797602

OPEN ACCESS

EDITED BY
Angel Chan,
Hong Kong Polytechnic University, Hong Kong SAR, China

REVIEWED BY
Mila Vulchanova,
Norwegian University of Science and Technology, Norway
Weifeng Han,
Federation University Australia, Australia

*CORRESPONDENCE
Jingdan Yang
jingdan.yang@uni-potsdam.de
Nan Xu Rattanasone
nan.xu@mq.edu.au

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

RECEIVED 19 October 2021
ACCEPTED 12 September 2022
PUBLISHED 14 October 2022

CITATION
Yang J, Kim J-H, Tuomainen O and Xu Rattanasone N (2022) Bilingual Mandarin-English preschoolers' spoken narrative skills and contributing factors: A remote online story-retell study. *Front. Psychol.* 13:797602. doi: 10.3389/fpsyg.2022.797602

Bilingual Mandarin-English preschoolers' spoken narrative skills and contributing factors: A remote online story-retell study

Jingdan Yang[1,2,3]*, Jae-Hyun Kim[4], Outi Tuomainen[1] and Nan Xu Rattanasone[4]*

[1]Department of Linguistics, University of Potsdam, Potsdam, Germany, [2]Faculty of Arts, University of Groningen, Groningen, Netherlands, [3]Philosophical Faculty, University of Eastern Finland, Joensuu, Finland, [4]Macquarie University Centre for Language Sciences, Multilingualism Research Centre, Department of Linguistics, Macquarie University, Sydney, NSW, Australia

This study examined the spoken narrative skills of a group of bilingual Mandarin–English speaking 3–6-year-olds (*N*=25) in Australia, using a remote online story-retell task. Bilingual preschoolers are an understudied population, especially those who are speaking typologically distinct languages such as Mandarin and English which have fewer structural overlaps compared to language pairs that are typologically closer, reducing cross-linguistic positive transfer. We examined these preschoolers' spoken narrative skills as measured by macrostructures (the global organization of a story) and microstructures (linguistic structures, e.g., total number of utterances, nouns, verbs, phrases, and modifiers) across and within each language, and how various factors such as age and language experiences contribute to individual variability. The results indicate that our bilingual preschoolers acquired spoken narrative skills similarly across their two languages, i.e., showing similar patterns of productivity for macrostructure and microstructure elements in both of their two languages. While chronological age was positively correlated with macrostructures in both languages (showing developmental effects), there were no significant correlations between measures of language experiences and the measures of spoken narrative skills (no effects for language input/output). The findings suggest that although these preschoolers acquire two typologically diverse languages in different learning environments, Mandarin at home with highly educated parents, and English at preschool, they displayed similar levels of oral narrative skills as far as these macro–/micro-structure measures are concerned. This study provides further evidence for the feasibility of remote online assessment of preschoolers' narrative skills.

KEYWORDS

narrative skills, Mandarin-English bilinguals, preschoolers, macrostructure, microstructure

Introduction

Children's early narrative abilities are important for their later literacy skills and play an important role in predicting their general academic performance as well as social and communicative success (Westerveld and Gillon, 2010; Gardner-Neblett and Iruka, 2015; Glisson, 2017; Pinto et al., 2017). Across different languages and cultures, narrative tasks are used as an ecologically valid way of collecting spoken language samples as they provide rich information about children's language abilities in a naturalistic setting (Botting, 2002; Boerma et al., 2016). For bilingual children, there is a paucity of evidence on the spoken narrative abilities especially for those speaking two typologically distinct languages, such as Mandarin and English. In the United States, Chinese languages (including Mandarin) are spoken by around 2.9 million people at home and are the most frequently spoken home languages other than English and Spanish (United States Census Bureau, 2021). In Canada, Mandarin is one of the most spoken home languages other than English and French (Statistics Canada, 2017). Similarly, in Australia, Mandarin is the most spoken home language other than English (Australian Bureau of Statistics, 2017). Despite the large number of bilingual Mandarin-English communities, little is known about the spoken narrative skills of these bilingual children in each of their two languages. This is especially the case for emerging bilingual preschoolers learning a home language (Mandarin) and a community language (English).

Recently, the COVID-19 pandemic has added to the challenges of testing young children and highlighted the need to move traditional face-to-face testing methods to remote online testing. There is emerging evidence to suggest that remote online testing of child language can be feasible, reliable, and valid (e.g., Sutherland et al., 2017; Manning et al., 2020; Sheng et al., 2021). In this study, the story retell task is used to assess bilingual preschoolers' spoken narrative skills in each of their two languages to address two aims: First, to add to our understanding on the spoken narrative skills of preschoolers learning two typologically distinct languages (Mandarin vs. English) and, second, to report on factors that predict bilingual preschoolers' performance on a remote online spoken narrative task.

to document the spoken narrative competence of a group of bilingual Mandarin-English preschoolers, to enrich our knowledge base on bilingual narrative competence in preschoolers learning two typologically diverse languages.

Spoken narrative skills

Spoken narrative skills, defined as the telling or retelling of a sequence of causally related events, requires the narrator to include detailed information about not only the setting, character, and theme of a story, but also the characters' actions, emotions, and motivations (Westby, 1991; Glisson, 2017). Spoken narrative skills are evaluated on levels of macrostructure and microstructure. Macrostructure refers to the global organization of a story, consisting of a "setting" plus one or more "episodes." The "setting" introduces the main character(s) and describes the context of the story (e.g., where the story takes place); and an "episode" includes an initiating event, the character's goal and attempt in response to the initiating event, and its consequences (Stein and Glenn, 1975; Gillam et al., 2016). Therefore, macrostructure requires adequate higher-level cognitive organization and abilities to conceptualize and plan sequences of events, as well as making inferences about the characters' motivations to convey a thematically coherent story (Bohnacker, 2016; Rezzonico et al., 2016).

Microstructure, on the other hand, relates to linguistic properties of the narrative in the target language (Stein, 1988; Squires et al., 2014; Bohnacker, 2016; Gillam et al., 2016). The evaluation of microstructure includes not only fine-grained linguistic structures used to construct a coherent narrative discourse, such as specific lexical and grammatical elements (Justice et al., 2010), but also more general measures about the overall spoken language productivity and syntactic complexity in the narrative genre (Westerveld and Gillon, 2010), e.g., total number of utterances, number of words, mean length of utterance (MLU), etc. Microstructures can therefore potentially provide a more detailed profile of a child's spoken language skills including their strengths and weaknesses in various spoken language domains of morphology, syntax, and semantics (Westerveld and Gillon, 2010).

One of the commonly used methods of eliciting spoken narratives from young children is through a story-retell task. Children are asked to first listen to a story and then reproduce the story, sometimes using visual support (Sheng et al., 2019). With the support of having listened to a prior story script, it is considered less demanding than other narrative tasks, such as story generation, in which children have to construct stories on their own. Therefore, the story-retell task is particularly appropriate for eliciting spoken narratives from younger preschool-aged children and bilinguals (Merritt and Liles, 1989; Westerveld and Gillon, 2010). Over and above the lower task demands, story-retell allows for experimental control over linguistic aspects such as length and complexity in the model story (Pearson, 2002).

Research on bilingual preschoolers

In terms of macrostructure, it has been suggested that its organization may be universal or invariant across languages (e.g., Berman and Slobin, 1994; Verhoeven and Strömqvist, 2001). Many studies report no differences in macrostructure measures between the two languages of bilingual children (Pearson, 2002; Squires et al., 2014; Bohmacker, 2016; Gagarina et al., 2016; Kunnari et al., 2016; Bonifacci et al., 2018; Méndez et al., 2018), especially for older school-aged children (Pesco and Bird, 2016). More recently, however, Hao et al. (2019) found that Mandarin-English bilingual preschool to school-aged children in the US performed better on "setting" in English than in Mandarin. This

could be due to English being the majority/community language leading to better performance compared to the home language (Pesco and Bird, 2016). However, the differences in scores for "setting" were small, suggesting that the macrostructure performance was, in general, still largely similar between the two languages (Hao et al., 2019). It is also unclear whether better English performance would be associated with age as school-aged children receive more formal education, including narrative skills, in English compared to preschoolers.

Microstructure, on the other hand, is more susceptible to variation across bilingual children's two languages (Pearson, 2002; Uccelli and Páez, 2007; Squires et al., 2014; Hipfner-Boucher et al., 2015; Boerma et al., 2016). This is not surprising given microstructures likely reflect differences in linguistic structures across languages. For example, Spanish-English-speaking 4-6-year-olds showed a strong association among microstructure elements within the same language, but more variation across languages, suggesting that these children are acquiring linguistic structures independently across the two languages (Méndez et al., 2018). On the other hand, in Hao et al. (2019) sample, while microstructure domains of "nominal" and "phrase" showed no significant differences between Mandarin and English, both "modifier" and "verb" were significantly better in English than in Mandarin. The pattern of performance on the various domains also differed within each language. For Mandarin, children were most productive in the "verb" and "nominal" domains, followed by "phrase" and "modifier," while in English, children produced more "verbs" than the other three domains. In general, these children demonstrated better narrative performance in English than Mandarin, but the differences were larger in *microstructure* than in *macrostructure*, further suggesting that macrostructure is less variable across languages than microstructure (Hao et al., 2019).

Language experience and bilingual narrative skills

One of the most important sources of influence on language acquisition, apart from general development, is language experience. Earlier age of acquisition and longer use typically lead to better language outcomes (Birdsong, 2009; Bosch et al., 2019; though see Xu Rattanasone et al. (2016) for different length of acquisition effects due to language typology in preschoolers). Bilingual children's language experience can also vary in terms of amount of language input and use, with both having effects on language development and spoken narrative skills (Hammer et al., 2012; Marchman et al., 2020). Govindarajan and Paradis (2019) found in school-aged children that length of English exposure in school predicted better English narrative skills, but amount of English input (from non-native speakers) and use at home did not predict macrostructure or microstructure abilities in English. Similarly, Hao et al. (2019) found that neither English input or output (production) correlated with performance on English macrostructure or microstructure narrative skills. Given the large age range in their study (4 to 9 years), and the lack of research on bilingual preschoolers' narrative skills, it is unclear whether these findings would specifically apply to preschoolers. Unlike school-aged children, preschoolers who have not yet received formal education in English (including explicit instructions on narrative skills), are cognitively and linguistically less developed, and therefore their narrative skills might be more influenced by different levels of language input.

The current study

Although Mandarin is one of the most common home languages around the world, only a few studies have examined the narratives skills of Mandarin-English bilingual children. While Hao et al. (2019) study provided a first important glimpse into the spoken narrative skills of these bilingual children, their study included a sample of children with a wide range of ages from preschool to primary school with varying lengths of exposure to English. This raises questions about the narrative skills of younger preschoolers. Such knowledge will provide us with more insights into early bilingual narrative development and could help inform educators on the language skills of typically developing bilingual preschoolers and their readiness for school.

This study examined a sample of bilingual 3–6-year-olds speaking Mandarin as their home language and learning English as the community language at childcare/preschool. In Australia, children of this age range typically attend a government subsidized private childcare (3–4-year-olds) or a fully funded government preschool (4–5-year-olds) or a kindergarten (5–6-year-olds) with English as the language of instruction. Their narrative skills were examined using a remote online story re-telling task in each of their two languages along with a weekly diary detailing daily Mandarin, English, and Mixed language input, and output for every awake hour. The entire study was conducted through remote online delivery. The following three research questions were addressed.

Research question 1

Whether there is a difference between bilingual preschoolers' macrostructure scores across their Mandarin and English; and whether there are any differences across macrostructure elements within each language. We predict a positive correlation between Mandarin and English, and no significant differences in overall performance on specific macrostructures across languages, but performance levels on macrostructure elements within languages might vary (Hao et al., 2019).

Research question 2

What are bilingual preschoolers' spoken narratives skills in terms of microstructures across their two languages? Is there a difference between bilingual preschoolers' microstructure domain scores in Mandarin and English? Within each language, are there

any differences in the production of individual microstructure domains? We predict that there could be no correlations cross-language in microstructure domains (unlike macrostructure; Hao et al., 2019). For Mandarin, performance should be best in the "verb" and "nominal" domains, followed by "phrase" and "modifier," and for English, performance should be best on "verbs" compared to the other three domains (Hao et al., 2019).

Research question 3

Are there any associations between preschoolers' narrative performance (macro/microstructural) and various contributing factors, such as chronological age, length of language exposure, and language input and output? We predict that chronological age would correlate positively with macrostructure (general developmental effect) and length of language exposure would correlate positively with microstructure (linguistic experience effect; Hao et al., 2019). We further predict that language input and output would correlate with performance in both languages (preschoolers are not yet receiving systematic schooling on narrative skills in English, unlike the school-aged children in the Hao et al., 2019).

Materials and methods

Participants

A total of 25 Mandarin-English (ME) bilingual preschoolers participated in this study but 5 were excluded for not switching languages, i.e., produced both stories only in English ($N = 1$) or only in Mandarin ($N = 2$), or could not finish either story ($N = 2$). The final sample consisted of 20 children (15 girls and 5 boys) aged between 3;10 (year; month) and 6;4 [mean age = 4;11, Standard Deviation (SD) = 8.7 months]. The primary carers of these children were in general well educated with two having received vocational training, eight completed an undergraduate degree and 10 postgraduate degrees. Of these primary carers 14 have received their highest level of training in Australia using English and 6 in China using English (2) or Mandarin (4). All participants resided in Sydney, Australia at the time of testing. Ethical approval for remote online testing was only allowed at the time and obtained from Macquarie University (approval number: 52021662724256).

Each child's primary caregiver was asked to complete a demographic information and language history questionnaire (see Appendix A). All children were raised in Mandarin-speaking households with native Mandarin-speaking parents (only one parent grew up speaking Cantonese and English). All parents were born in China (mainland or Hong Kong). The children's average age of acquisition (AoA) for English was 22 months (SD = 8.5 months; range: 11–41 months). All children were exposed to English through childcare before the age 36 months, except for one child at 41 months. The average length of English exposure was 37 months (SD = 12.7 months, range: 20–62 months). No language disorder or hearing impairment was reported.

Materials

Each child completed two story-retell tasks, one in English and one in Mandarin. Two different sets of wordless story pictures designed to elicit age-appropriate languages abilities in each language were used to avoid practice effects. The English picture story was *Ana gets lost* (Westerveld and Gillon, 2010), for children 4;0 to 7;6. The Mandarin story was taken from the "学龄前儿童语言能力测试 [Language Proficiency Test for Preschool Children]" [天津师范大学语言研究所 (Tianjin Normal University), 2016], for children 3;0 to 7;0. The stories were pre-recorded by a female native Australian English speaker and a native Mandarin speaker.

The primary caregiver of each child also provided a 7-day diary of hourly activities children were engaged in, interlocutors, if any, and language(s) heard (input) and produced (output) by the child. The diary data was later coded into total hours of hearing and speaking Mandarin, English, or both languages (mixed). The percentages of input and output of each language were then calculated by dividing the total number of hours hearing or speaking that language with the total number of awake hours.

The diary data is summarized in Table 1. Since the percentages of input to and output from children in each language were similar within languages, a single measure for each language was derived by averaging across input and output for that language. This new measure was used as a general indicator of language experience for each language. A paired *t*-test conducted between the average percentages of English and Mandarin language experience score found that children had significantly more experience in English

TABLE 1 Means, standard deviations (SD) and range for percent of input and output in each language.

Measure	English			Mandarin			Mixed		
	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Input	47%	13%	27–70%	30%	19%	2–58%	23%	19%	0–62%
Output	48%	14%	20–76%	27%	20%	1–58%	25%	20%	0–61%
Average	47%	13%	24–73%	29%	19%	2–58%	24%	19%	0–59%
t-test	$t = 3.11$, $p < 0.01$, Cohen's d = 0.70								

(Mean = 47%) than in Mandarin (Mean = 29%). The mixed language experience data (both input and output) was not included as we were only interested in pure English and Mandarin language experiences (mixed language would include both languages therefore masking the independent effect of each language).

Procedures

Elicitation

The data collection was done during the COVID19 pandemic, as a result, the testing took place *via* remote online delivery using Zoom. Children were invited to attend Zoom meetings with camera on and in the company of their parent(s). Bilingual Mandarin-English-speaking research assistants were trained to administer the tests *via* Zoom. Instructions were given to parents that they need to allow their child to complete the tests independently without any help, but they could encourage their child to pay attention to the task. It was also explained to each child and their parent that they were not expected to remember every detail of the story.

The tasks were administered first in Mandarin and then in English. The order of presentation was not counterbalanced as the study was conducted remotely online, in the children's homes, with help from their Mandarin-speaking parents and so all sessions began in Mandarin. Also, having their parents encourage them to participate in the home language helped ensure better engagement from the child participants. In the beginning of the story-retell task, children were instructed to carefully listen to a story. A set of wordless pictures appeared on screen one by one with the audio-recording of the story. After the story had finished, the same pictures were presented to the children again and they were asked to retell the story in their own words. If children did not start retelling the story spontaneously, prompts (e.g., "What happened in the beginning?") were used to help elicit responses. Parents were asked not to provide answers or repeat the answers. During each session, the instructions were given to children only in the target language. For children who could not complete the story-retell in one session, another Zoom session was arranged ($N = 5$).

Transcription

The recordings of children's story-retell were transcribed by a ME bilingual speaker (the first author) in ELAN (Max Planck Institute for Psycholinguistics, 2021) according to the CHAT transcription format (MacWhinney, 2019). Following the previous convention used in Hao et al. (2019), all task-related speech produced by the child (in forms of sentences, clauses, phrases, or single words) were segmented into communication units (C-unit), which is a main clause with its modifiers (Loban, 1976). Within each C-unit, transcription was done at the word-level for both Mandarin and English: Mandarin narratives were transcribed into written Chinese characters and English narratives into written English words. Verbal instructions from the experimenter, interventions from parents/caregiver, or task-unrelated speech or non-speech sounds (sighs, sneezes, coughs, crying, laughing, etc.) produced by the child were excluded from transcriptions. Inter-rater reliability was conducted between the first author and the last author (also a ME bilingual speaker) on 10% of the recordings in both Mandarin and English. Inter-rater agreement was 73.3% for C-units coded across the two raters (kappa = −0.134, $z = 0.974$, $p = 0.330$) suggesting substantial to high agreement (McHugh, 2012). On closer examination, the mean percent of disagreement across all C-units transcribed was 7.3% between the two raters (i.e., mean number of disagreements/number of agreements + number of disagreements per C-unit).

Coding and scoring

The evaluation included macrostructure and microstructure analyses. For macrostructure analysis in English, we chose the Story Quality Rubric (Westerveld and Gillon, 2010) as it was originally designed to analyse the macrostructure for "Ana gets lost." The decision was also made based on the consideration that our participants were younger than those in Hao et al. (2019) study, and their narrative productions were much simpler. Therefore, using other more complex rubrics such as Monitoring Indicators of Scholarly Language (MISL; Gillam et al., 2016) is likely to lead to overall poor performances (a floor effect). For macrostructure analysis in Mandarin, the rubric was adapted from the English version. Both macrostructure rubrics contained eight elements: *Introduction, Theme, Main Character, Supporting Character(s), Conflict, Coherence, Resolution and Conclusion.* Based on different levels of completion, each child was awarded different points for each element: 5 points if the child showed proficient ability in supplying the required details, 3 points if the child showed emerging ability in providing some details, and 1 point if the child provided minimal or no information. The scores were summed up to yield a total macrostructure score for each language. The possible minimum score was 8 and the maximum score 40. Details about the macrostructure scoring criteria for English and Mandarin can be found in Appendix B,C.

For microstructure analysis, we analysed the language samples for both general and fine-grained microstructures. Four general microstructure measures were evaluated to provide information about children's general narrative skills: total number of words (TNW), number of different words (NDW), total number of C-units (TNC), and MLU in words (MLUw). Data were extracted in CLAN (MacWhinney, 2000) with the "*freq*" and "*mlu -t%mor*" commands.

For the measures of fine-grained microstructures, we modified the Narrative Assessment Protocol (NAP) by Justice et al. (2010). Four domains of language (*phrase structure*, *modifier*, *nominal*, and *verb*) were evaluated, and each contained four to six elements. The English microstructure rubric contained 10 elements in four domains: *Phrase* (passive structure/locative phrase/temporal phrase); *Modifier* (adjective, adverb, negation); *Nominal* (personal pronoun), and *Verb* (copula/irregular past tense/regular past

tense). The rubric used for Mandarin microstructure analysis was an analogous rubric to the English version. Items that do not have analogous structures in Mandarin (e.g., English verb inflections) were excluded; unique features of Mandarin grammar (e.g., "ba" structure) were added. The final Mandarin rubric contained the same number of elements under four domains: *Phrase* ("ba" structure, locative phrase, temporal phrase); *Modifier* (adjective, adverb, classifier); *Nominal* (personal pronoun) and *Verb* (progressive aspect marker, perfective aspect marker, resultative aspect marker). However, unlike Hao et al. (2019), we did not include the Mandarin passive, "bei" structure, in our rubric. The Mandarin story we used, developed specifically for Mandarin speaking preschoolers, did not have passive structures. Indeed, Yip and Matthews (2007) showed that Cantonese-English simultaneous preschoolers (acquiring both languages from birth) did not produce 'bei' or passives in general with high frequency (Cantonese is a closely related Sinitic language to Mandarin). See Appendix D,E and for the English and Mandarin detailed microstructure rubrics.

Following Hao et al. (2019), for each microstructure element, we used both the 0–3 scale frequency score as in the NAP (Justice et al., 2010) and the raw frequency score. For 0–3 frequency score, each element was given a maximum score of three (even if the occurrence was more than three). The raw frequency considers all occurrences of an element and reflects children's productivity on that element. For example, if an element occurred 4 times, it was scored 3 for the 0–3 frequency score and 4 for the raw score. Consistent with NAP, for *Modifiers*, *Nominals*, and *Verbs*, only unique usages (types) were counted; but unique usage was not required for scoring *Phrasal* elements (tokens). Since there was only one unique personal pronoun in the Mandarin story, and the production of classifiers was limited, we scored tokens instead of types for these two elements. Only accurate productions of microstructure elements were counted.

Statistical analysis

To answer the first and the second research questions about differences in bilingual ME children's macro/microstructures between Mandarin and English, and among individual macrostructure elements/microstructure domains within each language, Linear mixed effects models were fitted using the lme4 package (Bates et al., 2015) in R (R Core Team, 2021). Following Hao et al. (2019), language experience was included in the models as a covariate. Mandarin language experience was subtracted from English, generating a difference score for each participant, which was then entered into the models as a covariate for both macrostructure and microstructure analyses (mean difference score = 19%; SD = 27%; range: −32 – 64%).

Regarding the fixed effect(s), for the *macrostructure* analysis, only "language" (Mandarin vs. English) was included in the model, however for the *microstructure* analysis, both "language" and "domain" (*phrase structure*, *modifier*, *nominal*, & *verb*) were included in the model. Varying intercepts were fitted for participants as random effects. The same linear mixed effects model was successively fitted for every macrostructure element/microstructure domain. For within language comparisons, pairwise *t*-tests with Bonferroni adjustments (Benjamini and Hochberg, 1995) were conducted between macrostructure elements/microstructure domains of each language.

The third research question was whether there were associations between children's narrative performance (macro/microstructural) and various contributing factors. Considering the relatively small sample size, the non-parametric Spearman correlation test was conducted on narrative performance scores (macrostructure English/macrostructure Mandarin/microstructure English/microstructure Mandarin, abbreviated as mac_eng / mac_man / mic_eng / mic_man respectively), chronological age (Age), age of acquisition for English (AoA), length of English exposure (E_length), language experiences (average of input and output) in English/Mandarin condition (E_Exp / M_Exp), and parent's bilingual dominance scores (Bilingdom; derived from the language history questionnaire). The Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) was conducted to avoid false discovery from multiple tests.

Results

Macrostructure

First, regarding cross-language comparisons, macrostructure total scores did not differ significantly, suggesting that overall performance across the two languages did not differ (Table 2). In terms of individual macrostructure elements, participants differed only on "Main character" and "Supporting character" across the two languages, with better performances in Mandarin than in English.

In terms of within-language comparisons, in both English and Mandarin, children scored higher on "Theme," "Main character," "Supporting character," "Resolution" and "Conclusion" than on "Introduction," "Conflict" and "Coherence" (see Appendix F for

TABLE 2 Parameter estimates for between language (Mandarin vs. English) comparisons on macrostructure.

Measure	English	Mandarin	*F*	*p*
	Mean (SD)	Mean (SD)		
Introduction	2.60 (1.67)	2.40 (1.60)	0.192	0.666
Theme	3.40 (1.90)	4.20 (1.01)	2.823	0.101
Main character	3.30 (1.63)	4.40 (1.31)	**6.066**	**0.024***
Supporting character(s)	3.10 (1.52)	4.20 (1.20)	**6.449**	**0.015***
Conflict	2.30 (1.49)	2.80 (1.82)	1.508	0.235
Coherence	2.50 (1.70)	2.40 (1.47)	0.045	0.834
Resolution	3.70 (1.49)	4.10 (1.02)	1.000	0.330
Conclusion	3.50 (1.28)	3.60 (0.94)	0.137	0.716
Total	24.40 (8.63)	28.10 (5.03)	4.135	0.056

Significant results are in bold. *$p < 0.05$.

pairwise *t*-tests), again showing similar patterns of performance across the two languages.

Microstructure

General microstructures

The general microstructure scores are presented in Table 3. The properties of the model stories are shown below under the "Model story" column. A "proportion" column indicates proportion of model-like structures produced by the children in relation to the target story. The Mandarin story was relatively simpler than the English story with fewer total C-units, total words, and different words, whereas the MLU in words (MLUw) was similar between the two languages. The descriptive data suggest that, as compared to the target story heard by the children, our participants produced relatively shorter narratives, with smaller numbers of TNC, TNW and NDW. They did however produce more model-like structures on each measure compared to the target story they heard for Mandarin than English.

Fine-grained microstructures

The 0–3 frequency scores and raw frequency scores of each microstructure element in two languages are summarized below in Table 4 (see Appendix G for proportion of model-like structures produced). The domain scores were derived by averaging across all elements within each domain, as shown in Table 4 (i.e., the mean and standard deviation of English Modifier elements across all participants were 1.75 and 1.12, respectively).

For mean 0–3 frequency scores in each domain, the results showed a significant main effect of domain [F (3, 57) = 32.48, $p < 0.001$], but they did not show a significant main effect of language [F (1, 133) = 2.74, $p = 0.100$] or interaction between language and domain [F (3, 133) = 1.24, $p = 0.298$]. Pairwise comparisons showed that all domains in English differed significantly with each other except between "Modifier" and "Verb" (Nominal > Modifier = Verb > Phrase), and all domains in Mandarin differed significantly with each other except between "Modifier" and "Nominal" domains (Nominal = Modifier > Verb > Phrase). See Appendix H for the pairwise comparisons.

Factors influencing narrative development

As shown in Table 5, no significant correlations were found between English and Mandarin for macrostructures or microstructures. Within each language, while there was a strong positive correlation between macrostructure and microstructure in English ($r_s = 0.799$, $p = 0.002$), macrostructure and microstructure did not reach significance for Mandarin.

Regarding potential contributing factors, chronological Age was significantly and positively correlated with macrostructures in both English ($r_s = 0.692$, $p = 0.012$) and Mandarin ($r_s = 0.711$, $p = 0.012$), but not with microstructures. English Age of Acquisition had no significant correlations with any narrative measure in either language. Length of English exposure, although showing a significant positive correlation with Age and negative correlation with Age of Acquisition, showed no significant correlation with either of the narrative measurements. Language experience in general did not significantly correlate with any narrative measures.

Discussion

Using remote online assessment, this study investigated Mandarin and English (ME) learning bilingual preschoolers' spoken narrative skills in terms of macrostructure and microstructure within and across languages. Potential factors influencing narrative development were also explored. ME bilingual preschoolers demonstrated similar narrative skills in their two languages, in both our measures of macrostructure and microstructure. More cross-linguistic differences were found in microstructure than macrostructure. Age was significantly positively correlated with macrostructures in both languages. Language experience, however, had no significant correlations with any aspects of children's narrative skills. These results were generally consistent with the findings from previous face-to-face studies (as outlined below), suggesting that it is feasible to use remote online spoken narrative tasks to measure bilingual preschoolers' spoken narrative skills.

TABLE 3 Means, standard deviations (SD), and ranges on general microstructure measures of child productions compared to model stories in English and Mandarin.

Measure	English					Mandarin				
	Mean	SD	Range	Model Story	Proportion%	Mean	SD	Range	Model Story	Proportion%
TNC	11.57	5.06	3–20	**24**	48.75	8.19	3.08	4–14	**12**	68.25
TNW	82.76	44.87	13–175	**193**	42.88	54.90	21.71	26–105	**118**	46.53
NDW	43.05	20.50	10–87	**115**	37.44	32.52	9.53	20–55	**63**	51.62
MLU-w	5.93	1.92	2.00–8.89	**8.04**		6.49	1.32	4.50–9.17	**9.83**	

The scores presented were averages of the two stories. TNC refers to total number of C-units; TNW refers to total number of words; NDW refers to number of different words; and MLUw refers to mean length of utterances in words. Significant results are in bold.

TABLE 4 Within language analysis of microstructure across domains for English and Mandarin in frequencies (0–3 and raw frequency).

Domain	English			Mandarin		
	Elements	Mean (SD)		Elements	Mean (SD)	
		0–3 frequency	Raw frequency		0–3 frequency	Raw frequency
Modifier	Adjective	2.05 (1.23)	2.45 (1.82)	Adjective	1.95 (1.00)	2.05 (1.19)
	Adverb	1.85 (0.99)	2.25 (1.62)	Adverb	2.10 (1.12)	2.45 (1.54)
	Negation	0.55 (0.60)	0.55 (0.60)	Classifier	1.90 (0.97)	1.90 (0.97)
	Average	**1.48 (0.74)**	**1.75 (1.12)**	**Average**	**1.98 (0.69)**	**2.13 (0.85)**
Nominal	Pronoun	2.15 (1.04)	2.80 (1.82)	Pronoun	2.30 (1.03)	4.00 (2.96)
	Average	**2.15 (1.04)**	**2.80 (1.82)**	**Average**	**2.30 (1.03)**	**4.00 (2.96)**
Phrase	Locative phrase	1.10 (1.21)	1.40 (1.82)	Locative phrase	1.45 (1.10)	1.45 (1.10)
	Passive phrase	0.35 (0.49)	0.35 (0.49)	Ba structure	0.60 (0.60)	0.60 (0.60)
	Temporal phrase	1.00 (1.08)	1.10 (1.29)	Temporal phrase	0.35 (0.49)	0.35 (0.49)
	Average	**0.82 (0.69)**	**0.95 (0.97)**	**Average**	**0.80 (0.45)**	**0.80 (0.45)**
Verb	Copula & Auxiliary	1.60 (1.19)	1.7 (1.38)	Perfective aspect marker	2.35 (0.81)	3.55 (2.16)
	Irregular past tense	2.15 (1.14)	2.80 (1.96)	Progressive aspect marker	0.55 (0.60)	0.55 (0.60)
	Regular past tense	1.15 (1.27)	1.60 (2.16)	Resultative aspect marker	2.15 (0.99)	2.55 (1.50)
	Average	**1.63 (1.05)**	**2.03 (1.68)**	**Average**	**1.68 (0.59)**	**2.22 (1.09)**
Total			**17.00(0.87)**			**19.45(1.25)**

Significant results are in bold.

TABLE 5 Correlation matrix of narrative skills with contributing factors.

	mac_eng	mic_eng	mac_man	mic_man	Age	AoA	E_length	Bilingdom	E_exp	M_exp
mac_eng		**0.799****	0.469	0.146	**0.692***	0.049	0.50000	0.471	−0.084	−0.399
mic_eng			0.416	0.379	0.5790	0.342	0.25000	0.302	0.105	−0.558
mac_man				0.518	**0.711***	0.073	0.40900	0.311	−0.264	−0.228
mic_man					0.2170	0.477	−0.14100	0.414	−0.346	0.079
Age						−0.078	0.734**	0.516	−0.075	−0.265
AoA							−0.690*0	−0.142	0.034	0.085
E_length								0.473	−0.145	−0.212
Bilingdom									−0.404	0.040
E_exper										−0.237
M_exper										

E: English; M: Mandarin; macro: macrostructure total score; micro: microstructure 0–3 frequency total score; Age: chronological age; AoA: English age of acquisition; E_length: length of English exposure; Bilingdom: parent's bilingual dominance score; E_Exp/M_exp: average percent of language input and output in English/Mandarin. *$p < 0.05$, **$p < 0.001$, indicates a significant correlation after the Benjamini–Hochberg procedure.

Bilingual narrative skills

Macrostructure

The first research question investigated whether there was any difference between bilingual ME preschoolers' macrostructure scores across languages, and whether there were any differences among various macrostructure elements within each language. We found that the overall performance on macrostructure measures did not differ between the two languages. Within both Mandarin and English, the preschoolers scored higher on "Theme," "Main character," "Supporting character," "Resolution," and "Conclusion" than on "Introduction," "Conflict," and "Coherence." These findings are generally consistent with previous studies (Bohnacker, 2016; Fiestas and Peña, 2004) and provide further support for children exhibiting different levels of ability across the different macrostructure elements at various stages of development.

However, there were differences in "Main character" and "Supporting character" elements between the two languages. For these two elements, children scored higher in Mandarin than in English. This is likely a reflection of the uneven complexity of the two stories used. In the Mandarin story, designed specifically for Mandarin-speaking preschoolers, the main character (the little rabbit) was introduced alone in the first sentence; while in the English story, the main character (Ana) was introduced along with two supporting characters (mom and dad). Regarding the supporting character(s), there was only one in the Mandarin story (the little hedgehog), whereas there were several in the English story (mom, dad, big brother Tom, & the policeman). Therefore, the cross-language differences that we observed in the "Main

character" and "Supporting character" are likely due to the differences of the two stories used to elicit narratives. Apart from these two elements, there were no significant cross-language differences on macrostructure, which aligns with previous findings that macrostructure reflects the global story organization and relies more on general cognitive abilities, so it tends to be less variable across languages (e.g., (Bohnacker, 2016; Rezzonico et al., 2016).

Microstructure

The second research question investigated any differences in bilingual ME preschoolers' microstructure scores between their Mandarin and English, and within each language. We predicted that microstructure was less likely to show cross-language similarities than macrostructure. But children might show different levels of ability on different microstructure domains within each language, due to the different linguistic characteristics of Mandarin and English. The results showed there were no significant interactions between language and domain, or on the main effect of language, indicating that our bilingual preschoolers performed similarly on their two languages. Our results differ from the commonly reported findings that microstructure usually differs across languages and is more variable than macrostructure (e.g., Pearson, 2002; Justice et al., 2010; Boerma et al., 2016). Indeed, Hao et al. (2019) showed different patterns for children's productions in Mandarin and English. However, it is possible that our sample of preschoolers have not yet developed enough linguistic competence or vocabulary in each language to show language-specific effects in spoken narrative skills. The different patterns across these two studies could indicate differences in stages of narrative development, and we had a much tighter sample of preschoolers who were predominantly 4- and 5-year-olds as opposed to the school-aged sample in Hao et al. (2019). The different results between our study and Hao et al. (2019) could also reflect difference in types of bilinguals. Our preschoolers were more akin to simultaneous bilinguals while Hao et al. (2019) sample of children had wider ages and ages of exposure to English, reflective of a mix of both simultaneous and early child L2 learners. Future studies should consider separating simultaneous and sequential bilinguals, as well as preschoolers from school aged children who are likely to be receiving structured reading and writing instructions in English (Whitehurst and Lonigan, 2001; Reese et al., 2010).

While our study did not indicate that our preschoolers were more dominant in English, Hao et al. (2019) study did show that their sample performed better in English than Mandarin, with greater language-specific knowledge in English. According to the authors, the significantly larger proportion of language experience in English (above 60%) than in Mandarin was responsible for driving the different performance across languages. Our study differs from Hao et al. (2019) in several ways. First, our participants were younger than those in Hao et al. (2019) study, who were exposed to English for longer with many receiving structured reading and writing instructions at school. Second, while our sample of preschoolers also had more language experience in English (mean = 47%) than in Mandarin (mean = 29%), the difference was not as large as in Hao et al. (2019) sample, i.e., 66–34% input and 75–25% output. It might be possible that smaller differences between home and community language experiences in our sample were not large enough to drive cross language differences in microstructures. Therefore, both developmental and language experiences might be driving different findings across the two studies which needs to be teased apart in any future research.

Factors influencing bilingual narrative development

The last research question relates to the effects of various factors that might contribute to ME preschoolers' narrative performances. Consistent with Hao et al. (2019), we found no direct relationship between language experiences (input and output) and macrostructure/microstructure in either Mandarin or English. This shows that despite some children receiving more English than Mandarin, this difference did not have any relationship with their spoken narrative skills in either language. The range in difference scores between Mandarin and English in our sample were not restricted in range and should be sufficiently large enough across the sample to show a relationship with spoken narrative skills if one exists: from −32% (more Mandarin) to +64% (more English). However, our sample also reported on average 20% Mixed language input and output (30% for Mandarin only and 50% for English only). Perhaps more needs to be understood about the role of mixed language output to better account for the language experience of these bilingual preschoolers, e.g., by using more intrusive and time intensive data collection methods such as audio and video recordings.

In terms of other factors, as predicted, older children in general showed better macrostructure performances than younger children in both languages; but age did not affect their microstructure performances. This suggests that macrostructure skills probably develop more rapidly during preschool than microstructure and is therefore easily detectable and affected by development. However, different from Hao et al. (2019), both age and length of English exposure failed to show any relationship with spoken narrative performance measures in either language. This might be due to restricted ranges in a sample of predominantly 4- and 5-year-olds.

Limitations and future directions for remote online testing

One limitation is that the English story was slightly more complex in macrostructure than the Mandarin story. Children in general supplied below 50% of the model story in C-units and word tokens and types for the English story, but between 50 and 60% for the Mandarin story. This might suggest that the English

story was too challenging for these bilingual preschoolers. In addition, given that MLUw was similar across the two stories and the patterns of performance were similar across the different microstructure domains and in both languages, the differences between the stories did not affect our purpose of making crosslinguistic comparisons. However, both the English and Mandarin stories lacked enough variety of certain microstructure elements, e.g., nominal items. As a result, only one element, personal pronouns, was included in the Nominal domain, making the overall microstructure evaluation less representative and comprehensive. The Multilingual Assessment Instrument for Narratives (MAIN), now with both Mandarin and English versions available, could be a good alternative all-in-one assessment for ME preschoolers (Gagarina et al., 2019; Luo et al., 2020).

Second, the language experience data collected from caregivers were inconsistent from child to child. Although we provided a template for documenting children's language activities and asked caregivers to provide as much detail as possible, some failed to conform to these instructions. Future studies should consider whether other methods of collecting real-time parental reports of diaries (e.g., sending text alerts) or samples of audio/video recordings might be more effective ways of collecting language experience data, albeit more intrusive and time intensive.

Third, the sample is relatively homogenous with limited age and ranges of English exposure. While this met the purpose of this study, i.e., examining bilingual preschoolers' spoken narrative skills, future studies should consider comparing preschoolers with school aged children or using longitudinal designs to capture developmental changes. Related to this, we did not examine the relationship between spoken narrative skills and individual variation on general cognitive abilities or combined vocabulary. This may be of interest for future studies to better understand individual variation (including sex) on macrostructures (global cognitive abilities) and microstructures (linguistic specific abilities) in dual language development. Future studies could also consider comparing similar children with their monolingual counterparts and school aged L2 learners to better understand effects of age of language acquisition (see Meisel, 2018). Such studies could also examine errors children make in each language to better understand the effect of language typology on dual language acquisition. For example, spoken Mandarin does not have gender pronouns (e.g., he/she) nor a case system (e.g., he/him). Mandarin is also an isolating language with no nominal (plurals) and verbal (tense) inflections. However, given that our sample is exposed to English as the community language very early in their development as preschoolers, we might expect this population to acquire structures in English similar to their monolingual peers (Meisel, 2004, 2009, 2018).

Finally, the global COVID19 pandemic has highlighted the need for more remote online research. While this study has provided some confidence in using remote online testing for assessing bilingual preschoolers' spoken narrative skills, more work is needed to evaluate the reliability and validity of remote online narrative elicitation tasks. One challenge we did experience is with managing parental intrusion, which was also reported in Du et al. (2020). When testing young children *via* remote online assessment, parent support is essential to enable the success of the session, however managing parental intrusion is much more difficult when parents must stay with their children to troubleshoot unexpected technical challenges. Our experience is that, for some parents, despite explicit instructions to encourage rather than provide linguistic support to their children, they continued to provide prompts despite explicit instructions at the beginning of each session and requests throughout the session to only provide encouragement. This would be especially disruptive in clinical settings. Such intrusions are also reported with similar sample of children (Du et al., 2020) and easier to manage in face-to-face testing where parents can wait in the same room but with some distance away from the child to avoid direct interference. Another challenge we experienced is reduced quality of some of the audio recordings due to not having control over the recording environment, e.g., children speaking too softly, poor internet connection, children who are unable to sit still, construction noise, etc. Coding these recorded productions offline can be challenging especially for untrained coders. While word level transcriptions can be made, acoustic analysis of segments could not be reliably conducted. One solution might be to send portable recorders to parents with detailed (written/video) instructions and the use of apps that indicate level of environmental noise (e.g., the ListenApp from Macquarie University).

Conclusion

Using remote online testing, this study investigated bilingual ME preschoolers' narrative skills at the levels of macrostructure and microstructure, both cross-linguistically and within-languages. The findings contribute to better understanding of the typical spoken narrative skills of ME preschoolers. It has provided further support that macrostructure develop similarly across the two languages, even during very early bilingual development in preschoolers. The age effect further suggests that cognitive development might be important in driving macrostructure development in narrative skills. The lack of effect of age of exposure to English for microstructures and lack of relationship between cross-linguistic experience on spoken narrative skills for ME speaking preschoolers may be an effect of typological distance between the two languages being acquired, but this needs further investigation. These findings add to our understanding of typical bilingual development and suggest that despite differences between home and community language experiences, it is possible

for children's dual language development in spoken narrative skills to be largely balanced.

Data availability statement

The raw data supporting the conclusions of this article will be made available upon request.

Ethics statement

The studies involving human participants were reviewed and approved by Macquarie University Humanities and Social Sciences (HASS) Human Research panel (approval number 52022662736215). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

JY wrote this thesis as part of her European Master's in Clinical Linguistics (EMCL+) and was responsible for transcribing the data, performing data analyses, interpretation of the results and drafting the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This project received funding from Macquarie University Centre for Language Sciences.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.797602/full#supplementary-material

References

Australian Bureau of Statistics. (2017). Cultural diversity in Australia, 2016. Available at: https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2071.0~2016~Main%20Features~Cultural%20Diversity%20Article~60 (Accessed January 17, 2022).

Bates, D., Maechler, M., Bolker, B. M., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.1017/S0142716415000399

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* 57, 289–300.

Berman, R., and Slobin, D. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum.

Birdsong, D. (2009). Age and the end state of second language acquisition. *The New Handbook of Second Language Acquisition* 17, 401–424.

Boerma, T., Leseman, P., Timmermeister, M., Wijnen, F., and Blom, E. (2016). Narrative abilities of monolingual and bilingual children with and without language impairment: implications for clinical practice: a narrative as diagnostic tool. *Int. J. Lang. Commun. Disord.* 51, 626–638. doi: 10.1111/1460-6984.12234

Bohnacker, U. (2016). Tell me a story in English or Swedish: Narrative production and comprehension in bilingual preschoolers and first graders. *Appl. Psycholinguistics* 37, 19–48. doi: 10.1017/S0142716415000405

Bonifacci, P., Barbieri, M., Tomassini, M., and Roch, M. (2018). In few words: linguistic gap but adequate narrative structure in preschool bilingual children*. *J. Child Lang.* 45, 120–147. doi: 10.1017/S0305000917000149

Bosch, S., Veríssimo, J., and Clahsen, H. (2019). Inflectional morphology in bilingual language processing: an age-of-acquisition study. *Lang. Acquis.* 26, 339–360. doi: 10.1080/10489223.2019.1570204

Botting, N. (2002). Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Lang. Teaching and Therapy* 18, 1–21. doi: 10.1191/0265659002ct224oa

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. 3rd Edn.* Mahwah, NJ: Lawrence Erlbaum Associates.

Du, Y., Sheng, L., and Salen, K. (2020, June). *"try your best" parent behaviors during administration of an online language assessment tool for bilingual mandarin-English children. In proceedings of the interaction design and children conference*, 409–420.

Fiestas, C. E., and Peña, E. D. (2004). Narrative discourse in bilingual children. *Lang. Speech Hear. Serv. Sch.* 35, 155–168. doi: 10.1044/0161-1461(2004/016)

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Bohnacker, U., et al. (2019). MAIN: Multilingual assessment instrument for narratives. Revised version. ZAS Papers in Linguistics 63:56. doi: 10.21248/zaspil.56.2019.414

Gagarina, N., Klop, D., Tsimpli, I. M., and Walters, J. (2016). Narrative abilities in bilingual children. *Appl. Psycholinguist.* 37, 11–17. doi: 10.1017/S0142716415000399

Gardner-Neblett, N., and Iruka, I. U. (2015). Oral narrative skills: explaining the language-emergent literacy link by race/ethnicity and SES. *Dev. Psychol.* 51, 889–904. doi: 10.1037/a0039274

Gillam, S. L., Gillam, R. B., Fargo, J. D., Olszewski, A., and Segura, H. (2016). Monitoring indicators of scholarly language: a Progress-monitoring instrument for measuring narrative discourse skills. *Commun. Disord. Q.* 38, 96–106. doi: 10.1177/1525740116651442

Glisson, L. (2017). A study to measure the efficacy of a manualised Oral narrative intervention Programme for school-age children with narrative delay [thesis, Curtin University].

Govindarajan, K., and Paradis, J. (2019). Narrative abilities of bilingual children with and without developmental language disorder (SLI): differentiation and the role of age and input factors. *J. Commun. Disord.* 77, 1–16. doi: 10.1016/j.jcomdis.2018.10.001

Hammer, C. S., Komaroff, E., Rodriguez, B. L., Lopez, L. M., Scarpino, S. E., and Goldstein, B. (2012). Predicting Spanish–English bilingual Children's language abilities. *J. Speech Lang. Hear. Res.* 55, 1251–1264. doi: 10.1044/1092-4388(2012/11-0016)

Hao, Y., Bedore, L. M., Sheng, L., and Peña, E. D. (2019). Narrative skills in two languages of mandarin–English bilingual children. *Int. J. Speech Lang. Pathol.* 21, 325–335. doi: 10.1080/17549507.2018.1444092

Hipfner-Boucher, K., Milburn, T., Weitzman, E., Greenberg, J., Pelletier, J., and Girolametto, L. (2015). Narrative abilities in subgroups of English language learners and monolingual peers. *Int. J. Biling.* 19, 677–692. doi: 10.1177/1367006914534330

Justice, L. M., Bowles, R., Pence, K., and Gosse, C. (2010). A scalable tool for assessing children's language abilities within a narrative context: the NAP (narrative assessment protocol). *Early Child. Res. Q.* 25, 218–234. doi: 10.1016/j.ecresq.2009.11.002

Kunnari, S., Välimaa, T., and Laukkanen-Nevala, P. (2016). Macrostructure in the narratives of monolingual Finnish and bilingual Finnish–Swedish children. *Appl. Psycholinguist.* 37, 123–144. doi: 10.1017/S0142716415000442

Loban, W. (1976). Language Development: Kindergarten through Grade Twelve. *NCTE Committee on Research Report No.* 18

Luo, J., Yang, W. C., Chan, A., Cheng, K., Kan, R., and Gagarina, N. (2020). The multilingual assessment instrument for narratives (MAIN): adding mandarin to MAIN. *ZAS Papers in Linguistics* 64, 159–162. doi: 10.21248/zaspil.64.2020.569

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. 3rd Edn*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Marchman, V. A., Bermúdez, V. N., Bang, J. Y., and Fernald, A. (2020). Off to a good start: early Spanish-language processing efficiency supports Spanish-and English-language outcomes at 4½ years in sequential bilinguals. *Dev. Sci.* 23:e12973. doi: 10.1111/desc.12973

Max Planck Institute for Psycholinguistics. (2021). ELAN| the language archive. ELAN. Available at: https://archive.mpi.nl/tla/elan (Accessed May 4, 2021).

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem. Med.* 22, 276–282. doi: 10.11613/BM.2012.031

Meisel, J. M. (2004). The bilingual child. *The handbook of bilingualism*, 91–113.

Meisel, J. M. (2009). Second language acquisition in early childhood. *Z. Sprachwiss.* 28, 5–34. doi: 10.1515/ZFSW.2009.002

Meisel, J. M. (2018). *Early child second language acquisition: French gender is German children*. Cambridge: Bilingualism: Language and Cognition,

Méndez, L. I., Perry, J., Holt, Y., Bian, H., and Fafulas, S. (2018). Same or different: narrative retells in bilingual Latino kindergarten children. *Biling. Res. J.* 41, 150–166. doi: 10.1080/15235882.2018.1456984

Merritt, D. D., and Liles, B. Z. (1989). Narrative analysis: Clinical applications of story generation and story retelling. *J. Speech Hear. Disorders* 54, 438–447.

Pearson, B. Z. (2002). "Narrative competence among monolingual and bilingual schoolchildren in Miami," in *Language and literacy in bilingual children* (United Kingdom: Multilingual Matters), 135–174.

Pesco, D., and Bird, E. K. R. (2016). Perspectives on bilingual children's narratives elicited with the multilingual assessment instrument for narratives. *Appl. Psycholinguist.* 37, 1–9. doi: 10.1017/S0142716415000387

Pinto, G., Bigozzi, L., Vezzani, C., and Tarchi, C. (2017). Emergent literacy and reading acquisition: a longitudinal study from kindergarten to primary school. *Eur. J. Psychol. Educ.* 32, 571–587. doi: 10.1007/s10212-016-0314-9

R Core Team (2021). R: A Language and environment for statistical computing. (Version 4.0) [Computer software]. Available at: https://cran.r-project.org

Reese, E., Suggate, S., Long, J., and Schaughency, E. (2010). Children's oral narrative and reading skills in the first 3 years of reading instruction. *Read. Writ.* 23, 627–644. doi: 10.1007/s11145-009-9175-9

Rezzonico, S., Goldberg, A., Mak, K. K.-Y., Yap, S., Milburn, T., Belletti, A., et al. (2016). Narratives in two languages: storytelling of bilingual Cantonese–English preschoolers. *J. Speech Lang. Hear. Res.* 59, 521–532. doi: 10.1044/2015_JSLHR-L-15-0052

Sheng, L., Shi, H., Wang, D., Hao, Y., and Zheng, L. (2019). Narrative production in mandarin-speaking children: effects of language ability and elicitation method. *J. Speech Lang. Hear. Res.* 63, 774–792. doi: 10.1044/2019_JSLHR-19-00087

Sheng, L., Wang, D., Walsh, C., Heisler, L., Li, X., and Su, P. L. (2021). The bilingual home language boost through the lens of the COVID-19 pandemic. *Front. Psychol.* 12:836. doi: 10.3389/fpsyg.2021.667836

Squires, K. E., Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., Bohman, T. M., and Gillam, R. B. (2014). Story retelling by bilingual children with language impairments and typically developing controls: story retelling by bilingual children with language impairments and typically developing controls. *Int. J. Lang. Commun. Disord.* 49, 60–74. doi: 10.1111/1460-6984.12044

Statistics Canada (2017). An increasingly diverse linguistic profile: Corrected data from the 2016 census. Available at: https://www150.statcan.gc.ca/n1/daily-quotidien/170817/dq170817a-eng.htm (Accessed January 17, 2022).

Stein, N. L. (1988). *The development of children's storytelling skill. In child language: A reader*. Oxford: Oxford University Press. (282–297).

Stein, N. L., and Glenn, C. G. (1975). *An analysis of story comprehension in elementary school children: A test of a schema*.

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

Uccelli, P., and Páez, M. M. (2007). Narrative and vocabulary development of bilingual children from kindergarten to first grade: developmental changes and associations among English and Spanish skills. *Lang. Speech Hear. Serv. Sch.* 38, 225–236. doi: 10.1044/0161-1461(2007/024)

United States Census Bureau. (2021). Detailed languages spoken at home and ability to speak English for the population 5 years and over: 2009–2013. Available at: https://www.census.gov/data/tables/2013/demo/2009-2013-lang-tables.html (Accessed January 17, 2022).

Verhoeven, L., and Strömqvist, S. (2001). *Narrative development in a multilingual context*. Amsterdam, The Netherlands: John Benjamins.

Westby, C. (1991). "Learning to talk, talking to learn: Oral-literate language differences," in *Communication skills and classroom success*, 334–357.

Westerveld, M. F., and Gillon, G. T. (2010). Profiling oral narrative ability in young school-aged children. *Int. J. Speech Lang. Pathol.* 12, 178–189. doi: 10.3109/17549500903194125

Whitehurst, G. J., and Lonigan, C. (2001). *Get ready to read! An early literacy manual: Screening tool, activities, & resources*. National Center for Learning Disabilities.

Xu Rattanasone, N., Davies, B., Schembri, T., Andronos, F., and Demuth, K. (2016). The iPad as a research tool for the understanding of English plurals by English, Chinese, and other L1 speaking 3-and 4-year-olds. *Front. Psychol.* 7:773. doi: 10.3389/fpsyg.2016.01773

Yip, V., and Matthews, S. (2007). Relative clauses in Cantonese-English bilingual children: typological challenges and processing motivations. *Stud. Second. Lang. Acquis.* 29, 277–300. doi: 10.1017/S0272263107070143

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 09 January 2023
DOI 10.3389/fpsyg.2022.832831

OPEN ACCESS

EDITED BY
Maria Teresa Guasti,
University of Milano-Bicocca, Italy

REVIEWED BY
Eleni Peristeri,
Aristotle University of Thessaloniki, Greece
Mike Putnam,
The Pennsylvania State University (PSU), United States

*CORRESPONDENCE
Natalia Gagarina
gagarina@leibniz-zas.de
Alexandra Perovic
a.perovic@ucl.ac.uk

†These authors share senior authorship

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Psychology

RECEIVED 10 December 2021
ACCEPTED 15 August 2022
PUBLISHED 09 January 2023

CITATION
Jažić I, Gagarina N and Perovic A (2023) Case marking is different in monolingual and heritage Bosnian in digitally elicited oral texts. *Front. Psychol.* 13:832831. doi: 10.3389/fpsyg.2022.832831

Case marking is different in monolingual and heritage Bosnian in digitally elicited oral texts

Ilma Jažić[1], Natalia Gagarina[2*†] and Alexandra Perovic[3*†]

[1]EMCL+, University of Groningen, Groningen, Netherlands, [2]Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany, [3]Department of Linguistics, Psychology and Language Sciences, University College London, London, United Kingdom

Heritage languages may differ from baseline languages spoken in the home country, particularly in the domains of vocabulary, morphosyntax and phonology. The success of acquiring and maintaining a heritage language may depend on a range of factors, from the age of acquisition of the second language; quantity and quality of input and frequency of first language use, to non-linguistic factors, such as Socio-Economic Status (SES). To investigate case marking accuracy in heritage Bosnian in relation to these very factors, we recruited 20 heritage Bosnian speakers in Austria and Germany, and 20 monolingual Bosnian speakers in Bosnia, aged between 18 and 30 years. Participants were assessed remotely in two sessions, on a battery of tests that included a background language questionnaire investigating participants' history of language acquisition, current usage and SES, and a newly adapted Bosnian version of the Multilingual Assessment Instrument for Narratives (MAIN). A significant difference in case marking accuracy was found between the two groups, despite the 97% correct performance in the heritage speakers, and an almost 100% performance of the monolinguals. In the heritage speakers group only, errors indicated a trend toward case system simplification as well as uncertainty in distinguishing between case meanings. The use of Bosnian, assessed through quantity and quality of input, as well as frequency of current usage, was shown to be a significant predictor of case marking accuracy in heritage speakers. In contrast, SES and age of acquisition of German did not play a role in these participants' case accuracy. The observed patterns of quantitative and qualitative differences in the case marking accuracy between heritage Bosnian speakers and their monolingual counterparts, in the face of a high level of accuracy, contribute to our understanding of the heritage language attainment in more diverse language dyads where L1 is a lesser studied language.

KEYWORDS

heritage language, bilingualism, case marking, narrative, nominal morphology, heritage grammars

Introduction

Bosnian is a morphologically complex Slavic language which is relatively under-researched, thus belonging to the category of lesser-studied heritage languages (Scontras and Putnam, 2020). Currently, Bosnia boasts a population of 3.5 million citizens. With 1.5 million speakers abroad, a considerable number of Bosnian speakers use and learn Bosnian within a bi- and multilingual context. As a result of the war in the 1990s, many Bosnian-speaking families immigrated to German-speaking countries (Gamlen, 2019). They have continued using Bosnian as their home language, transferring it to their children who now speak it as a heritage language.

The focus of our study is inflectional case morphology of young adult heritage Bosnian speakers in a Germanophone context. Case is defined as overt marking of the syntactic or semantic relationship of the noun with other elements within the same clause or sentence (Velupillai, 2012). The case marking of a noun is typically realized through affixes. Case marking can also be exhibited on adjectives, pronouns and determiners, however, for the purpose of this study, the focus will be on noun case marking. In this study, we investigate case morphology marking in adult heritage speakers, compared to monolingual speakers. We also consider various linguistic and non-linguistic factors that have previously been found to influence accuracy of morphosyntax in heritage languages: L2 age of onset (Anderson, 1999; Gagarina and Klassert, 2018), input and usage of the heritage language (Gathercole and Thomas, 2009; Kupisch, 2019; Czapka et al., 2021), and Socio-Economic Status (SES) (Sánchez, 1983; Cobo-Lewis et al., 2002).

Heritage bilingualism

The general consensus on the definition of a heritage language (HL) includes the following features: it is a minority language in a context of a majority language, HL speakers are bilingual and the majority language usually prevails as the dominant one in the adulthood (e.g., Lohndal et al., 2019). There are many aspects in which a HL may differ from the baseline/homeland language – the language as it is spoken in the home country: most notably in the domains of vocabulary, morphosyntax and pronunciation. In the domain of inflectional morphology, some cross-linguistic data point to a trend of rule simplification in the HL. Researchers argue that this language domain is particularly vulnerable to reanalysis of the underlying grammatical representation, a phenomenon also referred to as restructuring or, in some studies, variation (Montrul, 2015; Wiese et al., 2022). In terms of nominal morphology, this may be exhibited through a reduced or simplified case system, inconsistent use of gender, and subject–verb agreement errors. The simplification of the case system may also result in the omission of overt case markings (Montrul, 2015), thus resulting in a case system which reduces the opposition to nominative-accusative only in Russian, for instance (Polinsky, 2006, 2008). Other findings show oblique (non-nominative) cases in Russian heritage speaker production, however, their use is not always appropriate. For example, the loss of the differentiation between direction-location contrast, as expressed by the accusative and prepositional cases respectively (Isurin and Ivanova-Sullivan, 2008) as well as the use of nominative in the position of a direct object, the so-called *unification* of case features (Gagarina, 2017) has been observed.

On the other hand, a number of studies report contrasting results: that the heritage grammar shows no signs of simplification (Flores, 2015; Embick et al., 2020; Łyskawa and Nagy, 2020; Wiese et al., 2022). According to these authors, the variation found in the heritage grammar is a reflection of the variation that already occurs in the baseline grammar, differing only quantitatively – with heritage grammar having a higher incidence of variation. This discrepancy is attributed to the differences in the input received by monolingual and heritage speakers. The amount of language input available to heritage speakers is usually reduced compared to that of monolingual speakers (to be discussed below) and heritage speakers are more likely to receive input from the spoken register and/or non-standard variants. For instance, Łyskawa and Nagy (2020) found that case marking across heritage Polish, Ukrainian, and Russian was similar to that found in speakers of the languages spoken in these countries. Most variations observed in heritage languages were also noted in homeland languages (e.g., genitive-accusative substitution). The only exception was a default nominative assignment used solely in heritage languages. Case marking accuracy of heritage speakers has indeed been found to be robust, with the usual rates of accuracy reaching 90% and higher in different languages (Bolonyai, 2002; Hlavac, 2003; Rothweiler et al., 2010; Schmitt, 2010).

Child heritage speakers may demonstrate a slower rate of L1 case inflection acquisition compared to their monolingual counterparts. This involves a longer timeframe for developing case oppositions and uncertainty in determining the declension of nouns. Omitted and erroneous case marking forms are also observed in heritage speakers at an age when such a phenomenon is no longer found with monolingual children. Such a delay can occur if there is a considerable reduction in the amount of HL input upon L2 onset, as the case inflection may already be opaque and acquired relatively slowly even in a monolingual setting (Gagarina, 2011; Gagarina and Klassert, 2018). Additionally, some studies report a differential error pattern between structural and lexical cases in child heritage speakers. Structural case markings are more likely to be omitted, while lexical case markings show both omission and substitution errors (Bolonyai, 2002; Rothweiler et al., 2010).

10.3389/fpsyg.2022.832831

The role of linguistic and extralinguistic factors in heritage languages

Heritage speakers can either acquire the HL and the language of the environment simultaneously from birth or sequentially. In the latter case, the heritage speakers are raised in a monolingual HL environment until they enter the education system in the second language. This exposure usually occurs around the ages of 2 or 4, but it is not unusual for it to occur later, at the ages of 5 or 6. The age at which this exposure happens is referred to as the Age of Onset (AoO) or age of bilingualism (Kupisch, 2013). Distinctions are made not only between simultaneous and sequential bilinguals but also between different AoO groups within the sequential bilingual group. The reason these distinctions are made is because of the assumption that there exist multiple sensitive or critical age periods. Exposure to sufficient language input during these periods ensures a more successful acquisition of certain linguistic features. After these periods are complete, native-level attainment of those features within the first language timing and path is less likely. Informed by the findings on neurological development as well as the typical schedule of language acquisition, the proposed critical periods are ages 4–6 and ages 6–7 (Meisel, 2009, 2011). The onset of the L2 implies a decrease in the amount of HL input. This in turn may affect the level of success with which certain HL features are acquired or trigger attrition of already acquired HL features (Montrul, 2015).

The effect of the AoO of the L2 on the development and outcomes of heritage grammars has been widely investigated.[1] It has been demonstrated in different linguistic domains, from phonology to morphosyntax (e.g., Flege et al., 1999). Some studies argue for a sequential bilingual advantage in HL over simultaneous bilinguals. This is ascribed to a longer HL monolingual period and a later AoO of the society language (SL). This effect of AoO was found in HL domains such as gender agreement (Anderson, 1999) and aspectual contrasts in Spanish (Montrul, 2002) as well as case inflection and expressive lexicon in Russian (Gagarina and Klassert, 2018). However, some domains of HL grammar fail to show an effect from AoO: definiteness in Turkish (Kupisch et al., 2016), sentential negation and wh-questions in Greek (Makrodimitris and Schulz, 2021) and verb inflections in Russian (Gagarina and Klassert, 2018).

Heritage language input and use are thus crucial in the investigations of HL development. Both are complex, multidimensional concepts which need to be carefully dissected. There are a multitude of possible sources of linguistic/HL input such as from family and peers, educational institutions (school, preschool, day care) as well as media (books, TV, music) (Unsworth, 2016). It is useful to consider both the quantity and quality of input and use (Kupisch, 2019). Quantity can be inferred from the number of people (parents, siblings, friends) speaking the HL, the number of visits to the country of HL and activities carried out in HL (Kupisch, 2019). Quality of HL input is commonly gauged by the linguistic richness of the input and contextual diversity of HL exposure. The HL may be spoken by individuals whose language is rich and of standard variety or has already undergone attrition; the HL can be exclusively spoken or also written; it can be exclusively informal or it can be provided in educational contexts (Kupisch, 2019; see also e.g., Unsworth, 2015, 2016).

The effect of HL input and use on the development of the HL in children has been shown to influence the speed and manner of acquisition across different linguistic domains such as vocabulary, morphosyntax and semantics (Thomas and Gathercole, 2005; Gathercole and Thomas, 2009; Paradis et al., 2011; La Morgia, 2015; Montrul, 2015; Unsworth, 2015, 2016; Gagarina and Klassert, 2018; Kupisch, 2019; Czapka et al., 2021; Makrodimitris and Schulz, 2021). Variation in the quantity and quality of input as discussed above is considered by some the fundamental determinant of the interindividual variation observed in bilingual language acquisition (Paradis, 2011).

The quantity of HL input is known to affect vocabulary size as well as diversity of produced morphemes: children receiving more input are reported to perform better than those with less input (Gathercole and Thomas, 2009; Unsworth, 2015, 2016; Czapka et al., 2021). Importantly, Gagarina and Klassert (2018) and Makrodimitris and Schulz (2021) report the use of the HL at home to be a significant predictor for the grammar domains under investigation in their respective studies. In their study of local and distant gender marking in Welsh with Welsh-English bilingual children, Thomas and Gathercole (2005) found the amount of input to influence speed of acquisition, especially with regards to the more complex and less transparent structures (such as possessive *ei* for masculine nouns in Welsh). Such structures are acquired later: a lower amount of HL input would not suffice in ensuring the successful attainment of the feature during the critical period for its acquisition. There is further evidence that the input received during childhood, as well as throughout life, is critical for the development and maintenance of the HL in adulthood. Another study by Gathercole and Thomas (2009) found that the vocabulary levels of adult heritage Welsh speakers were affected by both the input from their childhood as well as the consistency of input they received as adults (e.g., language of the partner). The amount of HL input and use is often related to the status that the language enjoys in the social environment of the heritage speaker. The social value attributed to the HL will determine whether the

1 Here we do not refer to adult L2 learners: we assume that heritage grammars are different from L2 grammars in crucial ways, i.e., AoO is early rather than late and mode of instruction is naturalistic, not formal. See studies that report qualitative differences in the underlying linguistic knowledge of these populations for more information, e.g., Van Osch et al., 2018.

country's policies allow for education in that language or how present the language is in the public sphere in general, all of which ultimately affects the success with which it is mastered (Montrul, 2015). In Gathercole and Thomas (2009), the authors are mindful of the fact that the Welsh-English community is quite stable and large, which is not usually the case for immigrant bilingual communities.

With regards to non-linguistic factors, the role of SES (most often measured *via* variables such as education, income, and occupational prestige) in language development in monolingual contexts is reported to be vital. A higher SES is known to correlate with more advanced lexical and grammatical skills (Hoff, 2006), where quantity and quality of language input, amongst other factors, is argued to be particularly relevant in early lexical development (Hart and Risley, 1995). As for heritage speakers, and especially adults, the relationship between HL development and SES is less well understood. Lower SES Spanish heritage speakers in the US were found to use more HL daily and achieve higher oral proficiency compared to their higher SES counterparts (Sánchez, 1983; Cobo-Lewis et al., 2002). On the other hand, in the study of Armon-Lotem et al. (2011) of Russian–Hebrew and Russian–German speakers, no effect of SES was reported on the L1 maintenance for the Russian–Hebrew cohort, but was present in the Russian–German cohort. The authors explain the lack of an SES effect in the Russian–Hebrew group by the SES homogeneity of that particular group.

In sum, while some factors such as AoO, the prestige of the home language, or SES have attracted more attention in the literature, the role of other factors, such as quantity, and especially quality of input and use in heritage languages, are less well researched and understood.

Bosnian as a heritage language

Basic characteristics of Bosnian

A south Slavic language, Bosnian shares many properties with other Slavic languages, such as rich morphology, relatively free word order and a lack of articles. Number, case and gender markings are fused into a single suffix and are marked on all nominal elements: nouns, pronouns, adjectives and some numerals. Additionally, all of the nominal elements within an noun phrase (NP) express number, case and gender agreement. Verbs can be inflected for person, tense, aspect and mood, while subject-verb agreement includes features of number, person, and gender.[2] The sentence in the example (1) illustrates most of the characteristics above.

(1) Ona popravlja moju staru mašinu.
She $\text{repair}_{3SG.PRS}$ $\text{my}_{ACC.F.SG}$ $\text{old}_{ACC.F.SG}$ $\text{machine}_{ACC.F.SG}$
"She is repairing my old machine."

Case morphology in Bosnian and its acquisition

The Bosnian case system differentiates between seven cases (see **Table 1**). Based on the class of the noun, there are three basic types of declensions. The first one consists of masculine (not ending in -a) and neuter nouns, the second contains nouns ending in -a (feminine and masculine), while the third declension solely accepts feminine nouns with a zero ending.[3]

A pertinent phenomenon that occurs in case paradigms is syncretism – where distinct cases share the same form (see **Table 2** for examples relevant to Bosnian noun declensions). The nominative case is syncretic with the accusative case in the paradigms for inanimate masculine nouns, all neuter nouns

2 For the purpose of the discussions below, the relationship of Bosnian with Croatian, Montenegrin and Serbian must be noted. These languages until recently comprised a single language, Serbo-Croatian, which was the official language of Yugoslavia. However as Alexander (2006) remarks, it was always a "pluricentric" (p. xviii) language, which recognized several standard idioms. Following the dissolution of Yugoslavia, these distinct standard idioms came to officially form four separate languages: Bosnian, Croatian, Montenegrin, and Serbian (often investigated jointly under the umbrella of BCMS studies). Relevant findings from any of these languages will be included in our literature review.

3 Within these three main classes of declension, there are many sub-declensions based on the number of syllables and different phonological conditions which will not be discussed here (Alexander, 2006).

TABLE 1 Cases in Bosnian and their prototypical function and meaning.

Case	Function	Meaning
Nominative	Subject	Labeling
Accusative	Direct object	Object/goal (with prepositions)
Genitive	Possessor, missing entity, genitivus partitivus	
Dative	Indirect object	Recipient/goal
Vocative	Addressing someone	
Instrumental	Device or company	Means and company
Locative	Prepositional phrase – verb complement	Topic and location

TABLE 2 Examples of Bosnian noun declension, for Masculine, Neuter and Feminine genders, singular and plural forms: "konj" horse; "dan" day; "selo" village; "ruka" hand; and "stvar" thing.

	Masculine		Neuter	Feminine	
	Animate	Inanimate		-A ending	Consonant ending
N sg	kȍnj	dân	sèlo	rúka	stvâr
G sg	kònja	dâna	sèla	rúkē	stvâri
D sg	kònju	dânu	sèlu	rúci	stvâri
A sg	kònja	dân	sèlo	rúku	stvâr
V sg	kȍnju	dâne	selo	rúko	stvâri
I sg	kònjem	dânom	sèlom	rúkōm	stvâri
L sg	kònju	dânu	sèlu	rúci	stvâri
N pl	kònji	dâni	sela	rûke	stvâri
G pl	kònjā	dânā	sèla	rȕkū	stvârī
D pl	kònjima	dânima	selima	rúkama	stvârima
A pl	kònje	dâne	sela	rúke	stvâri
V pl	kȍnji	dâni	sela	rúke	stvâri
I pl	kònjima	dânima	selima	rúkama	stvârima
L pl	kònjima	dânima	selima	rúkama	stvârima

N, nominative; G, genitive; D, dative; A, accusative; V, vocative; I, instrumental; L, locative; sg, singular; pl, plural.

and feminine nouns with a zero ending. Therefore, in all of these paradigms, both the nominative case and the accusative case forms have a zero ending. For animate masculine nouns, the genitive case is syncretic with the accusative case. All paradigms also have syncretic forms for plural dative, locative and instrumental forms.

In order to better understand the properties of heritage Bosnian, here we provide a brief overview of monolingual child acquisition of case, in view of similarities between heritage speakers and child L1 learners (Polinsky and Scontras, 2020). The acquisition of nominal morphology and case in Bosnian children is not well documented, however, some evidence from Croatian does exist: the two languages are close enough to expect similar acquisition patterns in this domain of grammar. Before their second birthday, Croatian-speaking children already develop certain mini-paradigms (Kovačević et al., 2009). These paradigms are usually found for feminine nouns whose case forms are less syncretic. As such, they provide a clearer juxtaposition between the case markings in the input, which the children then utilize to construct mini-paradigms, usually contrasting 3–4 cases. All case markings emerge before age 1;10, with accusative markings first appearing at age 1;4, while locative and instrumental markings are among the last to occur at ages 1;9 and 1;10, respectively. At age 2;5, the distribution of cases already closely resembles that of the adult input language (Kovačević et al., 2009). The development of fully fledged paradigms for all lexemes in the child's mental lexicon is, however, a long and complex process – case morphology is characterized by non-transparency and syncreticity cross-linguistically which influence the rate at which it is acquired (Xanthos et al., 2011). Incorrect case forms of certain nouns can be found in pre-school as well as school age (Kovačević et al., 2009; Vrsaljko and Paleka, 2018). A common error of using the locative[4] (used to signify location) instead of accusative case (used to signify direction) appears in 2-year-olds: "i onda smo išli na placu [*]" (Hržica and Peretić, 2015), and is seen even later, at age 6: "...dok je on išao u krevetu [*]" (Hržica and Lice, 2013). If case poses a challenge for monolingual L1 acquisition, it can serve as a valuable foundation for making predictions on the outcomes of heritage language acquisition (Polinsky, 2018; Polinsky and Scontras, 2020). We could thus expect heritage speakers to diverge from standard usage of Bosnian with case markers that seem most problematic for Croatian child language L1 speakers, e.g., accusative with nouns signifying direction.

Heritage Bosnian

There are a handful of studies on heritage Bosnian, though mostly in an English-speaking context and amalgamated with the closely related Croatian and Serbian into heritage BCMS (Bosnian, Croatian, Montenegrin, Serbian) studies. While the majority of these studies take on sociolinguistic issues, such as quality of education in the heritage language or attitudes toward the HL (e.g., Ćatibušić, 2019), Hlavac (2003) focuses on the morphology of heritage Croatian speakers. This study included 100 participants aged 16–32 who were either born in Australia or moved there before the age of 5 with parents who originated from either Croatia or Bosnia and Herzegovina. The corpus created consisted of 15–20 min of transcribed speech segments of answers to open-ended questions as well as descriptions of visual stimuli. Some information on linguistic background, such as order of acquisition and use of HL with friends and family were gathered through a structured questionnaire, but no information on the level of education or other SES factors were provided. The recorded background linguistic factors were not included into the analyses as potential explanatory variables.

Heritage speakers used target case marking in more than 90% of cases. An example of non-target case marking is given in (2): the noun *rodbina* "extended family" is used erroneously in the unmarked nominative form instead of the overtly marked accusative form *rodbinu*. In their examples of intra-word code-switching, participants sometimes used an appropriate Bosnian case marker on the English NP (example 3). In some instances, however, an NP contained an unintegrated, directly transferred English noun as its head (example 4). In such cases, the

4 In addition to the locative case, some prepositions denoting location assign instrumental: e.g., "nad gradom" (above the city). We shall therefore refer to these as 'location or instrumental,' even when giving examples of prepositions assigning locative case only, as these seem more common in the sources we consulted.

10.3389/fpsyg.2022.832831

rest of the NP constituents which are congruent with the head noun, such as attributives and determiners, had a higher incidence of non-target markings. Thus in example 4, the preposition *na* "on" requires the locative case, but due to the unintegrated noun "side," the dependent attributive "other" is in the unmarked nominative case. The example in (5) illustrates the case and number mismatch found in heritage Croatian NPs: the possessive *njegov* "his" is singular and nominative, while the head *prijatelje* "friends" is plural and accusative.

(2) . . . I tu imamo rodbina*
and here have$_{1\,PL.PRS}$
extended family$_{NOM.F.SG}$
. . . "and here we have extended family"

(3) . . . I tamo sam dobio posao
and there be$_{1\,PSG.AUX}$ got$_{3SG.M.PTCP}$ job$_{ACC.M.SG}$
u hospital-u za treću godinu*
in hospital$_{LOC.M.SG}$ for third year$_{GEN.F.SG}$
. . . "and there I got a job in a hospital for a third year"

(4). . . . Gdje je plaža,
Where be$_{3PSG}$ beach
na drugi *side* ima . . .*
on$_{(+LOC)}$ other$_{NOM.M.SG}$ side have
". . .where the beach is, on the other side there is. . ."

(5) Trebam vodit moj
Must$_{1SG.PRS}$ take$_{INF}$ my$_{NOM.M.SG.}$
brat i njegov prijatelje*
brother$_{NOM.M.SG}$ and his$_{NOM.M.SG.}$ friend$_{ACC.M.PL}$
"I must take my brother and his friends"

Research on heritage BCMS in the German-speaking context have also had a sociolinguistic focus, especially on the issue of cultural identity (Savić, 1989; Schlund, 2006; Randjelovic, 2019). The main insights into the language features of heritage BCMS speakers in a Germanophone context comes from studies by Hansen et al. (2013) and Hansen (2018)[5]. These authors investigated heritage BCMS speakers aged between 20 and 32. Participants were either born in Germany or moved there before the age of five from either Bosnia and Herzegovina, Croatia, or Serbia. The corpus consisted of qualitative interviews in heritage BCMS from 11 participants, supported by written production data (essays written in a heritage language class) and speech samples elicited on the basis of four pictures. In line with Hlavac (2003), case incongruity between head nouns and their dependents was also observed (example 6). These BCMS heritage speakers seem to pattern with American Russian heritage speakers regarding the difficulty observed in dealing with the two-way prepositions assigning either accusative or locative (Isurin and Ivanova-Sullivan, 2008). In example (7) both cases are used, incurring another instance of case mismatch within an NP.

(6) . . . I kod nas su
and at we$_{GEN}$ be$_{3\,PL.PRS}$
one turski krovove*
this$_{ACC.M.PL}$ Turkish$_{NOM.M.PL}$ roof$_{ACC.M.PL}$
"And we have those Turkish roofs"

(7) Onda kaže na jednoj ruku
then say$_{3SG.PRS}$ on one$_{LOC.F.SG}$ hand$_{ACC.F.SG}$
isto što nisam
same what NEG·1SG.AUX
bio uvek
be$_{3SG.M.PST}$ always
"Then she says the same on the one hand that I was never there."

There are several instances that indicate transfer of a German argument structure, resulting in incorrect case marking. In example (8), the heritage speaker uses the preposition *protiv* "against" with an accusative, which is a structure corresponding to the German counterpart "gegen" but is erroneous as the BCMS preposition assigns the genitive. Similarly, in the heritage BCMS sentence (9), the existential verb *ima* "have" assigns the accusative, as is the norm in German, but this takes the nominative case in BCMS[6]. Heritage speakers in this corpus also exhibited deviations in gender agreement between nouns and their determiners, as observed in example (10).

(8) i dobili jednu jednu
and get$_{1\,PL.M.PTCP}$ one$_{ACC.F.SG}$ one$_{ACC.F.SG}$
utakmicu i. . . protiv
match$_{ACC.F.SG}$ and . . . against(+ GEN)
Mađare* dva-dva odigrali
Hungarian$_{ACC.PL}$ 2:2 play$_{1\,PL.M.PTCP}$
"We had one match against the Hungarians, we played 2:2."

(9) tamo u. ima njemačku
There at have$_{3SG.PRS}$ German$_{ACC.F.SG}$
poštu ili Telekom u Zagrebu*
post-office$_{ACC.F.SG}$ or Telekom in Zagreb$_{LOC.SG}$
"There at. in Zagreb there is a German post office or Telekom."

5 Findings from the Raecke (2006) corpus will not be discussed here, as they primarily discuss the use of clitics in Croatian heritage speakers.

6 In the existential sentences with the verb 'ima,' singular noun phrases are assigned nominative case, but plural nouns are assigned genitive, see e.g., Hartmann and Milicevic (2009) for discussion.

(10) i na taj vreme šta
and on this$_{ACC.M.SG}$ time$_{ACC.N.SG}$ what
da uspijem?*
COMP manage$_{1SG.PRS}$
"What can I manage in this time?"

Based on the studies above, the general features of heritage Bosnian have been documented and outlined. Building on this foundation, this study aims to provide a more precise picture of heritage Bosnian nominal morphology and to explain its variation.

The present study

The current study investigates case marking in adult heritage Bosnian speakers having grown up with German as their societal language, compared to adult Bosnian monolinguals. The first research question asks whether case marking accuracy differs between heritage speakers and monolinguals in obligatory contexts in elicited narratives. Based on previous findings from heritage Slavic languages, we predict, (a) lower accuracy in case marking in heritage Bosnian compared to the monolingual/baseline language; and (b) a restructuring of the heritage case system. By restructuring we mean the omission of overt case markings and reduction of case oppositions.

Our second research question is concerned with factors that influence the accuracy of case marking in the heritage speakers. Previous studies have emphasized the role of both linguistic (e.g., use and input) and extralinguistic factors (e.g., SES) in the development and maintenance of language proficiency as well as discrete features of grammar in children (Armon-Lotem et al., 2011; Gagarina and Klassert, 2018; Makrodimitris and Schulz, 2021). It is not well established whether these effects persist into adulthood for heritage speakers. We therefore ask whether the usage of heritage Bosnian, age of German onset and participants' SES predict case marking accuracy in narratives of adult Bosnian heritage speakers.

Materials and methods

Participants

We recruited two groups of participants aged between 18 and 30: 20 adult Bosnian heritage speakers (15 female) from Germany and Austria and a control group of 20 adult native Bosnian speakers (11 female) from Bosnia (see **Table 3**). All of the participants were healthy adults, with no neurological conditions and normal or corrected-to-normal vision. The heritage speakers had at least one Bosnian parent. They grew up in a German-speaking country or moved to one before the age of four. All of the participants in the control group were monolingual speakers who were born and had lived all of their lives in Bosnia and Herzegovina. They came from monolingual households, though a majority of them learnt at least one foreign language during their primary and secondary education. None had a university-level degree in a foreign language.

Background measures

Information about the participants' history of language acquisition and current usage was collected through a Language Background questionnaire, adapted to Bosnian from Lloyd-Smith (2020). Using the information collected through the questionnaire and adapting the procedure outlined in Lloyd-Smith (2020), a Bosnian Use Score for heritage speakers was calculated. This score quantifies heritage language use by considering four core aspects: Language Use at Home, Quality of Language Use, Current Language Use and Time Spent in Heritage Country. A detailed explanation of the weighted score calculation can be found in **Table 4**. The maximum possible score is 28.5, and for the current group of participants the mean score was 17.56 (SD = 3.03).

Additional background measures included Years of Education, as a measure of SES, and AoO for German. For the monolingual group, the mean number of years of education was 15.3 (SD = 2.36), not statistically significantly different (p = 0.637) from that of the heritage speaker group, whose mean number of years of education was 14.92 (SD = 2.62). The mean AoO of German for the heritage group was 3.8 (SD = 1.60).

Procedure and scoring

The data collection took place over the course of the summer of 2021. The test battery consisted of the aforementioned Language Background Questionnaire and the newly adapted Bosnian version of the Multilingual Assessment Instrument for Narratives (MAIN) (Gagarina et al., 2012, 2019b). The data was collected remotely using the conferencing software Zoom.

For the heritage speakers, there were two sessions per participant, lasting roughly 15 min each. One session consisted of filling out the Language Background Questionnaire followed by a MAIN assessment in Bosnian administered by the first author. The second session consisted of a MAIN assessment in German administered by a student research assistant working for the ZAS, Berlin, Germany. Monolingual participants were administered the MAIN and the Language Background Questionnaire in the course of one session. The counterbalancing for language and story order followed the procedures outlined in the MAIN manual (Gagarina et al., 2012, 2019a).

TABLE 3 Participant information.

Group	*N*	Age – Years (*SD*)	Years of education (*SD*)	Age of L2 onset (*SD*)
Monolingual	20 (11 females)	24.05 (3.06)	15.3 (2.36)	
Heritage speakers	20 (15 females)	23.35 (2.99)	14.92 (2.62)	3.8 (1.6)

TABLE 4 Bosnian usage score calculations.

Types of use	Scoring
Language use at home	
L of father L with father L of mother L with mother L siblings L grandparent L partner Distant relatives L at home (before age 6) L at home (after age 6)	1 pt = Bosnian 0.5 pts = Bosnian and German 0 pts = German
Quality of language use	
Number of years schooling in Bosnian	6 + years = 3 pts; 3 + years = 2 pts; 1–2.9 years = 1 pt; 0 year = 0 pts
Bosnian studies at the University	Yes = 1 pt; No = 0 pts
Number of Bosnian University courses	7–9 = 1.5 pts; 4–6 = 1 pt; 1–3 = 0.5; 0 = 0 pts
Number of contact types with Bosnian	Listening/speaking/reading/writing = 3 pts; 1 of 4 missing = 2 pts; 2 of 4 missing = 1 pt
Long period of Bosnian non-use	No = 1pt; Yes = 0 pt
Current Language Use	
Relative use of Bosnian vs. German	Bosnian 100% = 3 pts; 75% = 2.5 pts; 50% = 2 pts; 25% = 1 pt; 0% = 0 pts
L at work/school L at home L with friends	1 pt = Bosnian 0.5 pts = Bosnian and German 0 pts = German
Time spent in heritage country	
Number of years spent in Bosnia	2.1 + years = 3 pts; 1.1–2 years = 2.5 pts; 6.5–12 months = 2 pts; 4–6 months = 1.5 pts; 3–4 months = 1 pt; 1–2 months = 0.5 pts
Number of visits in past 5 years	7 + = 2 pts; 4–6 = 1.5 pts; 1–3 OR every other year = 1 pt; 0 = 0 pts

The procedure of the MAIN administration in Bosnian included the elicitation of two stories, "Cat" and "Baby Birds," through the telling mode. The order of the stories was counterbalanced with half of the participants telling the Cat story first, while the other half told the Baby Birds story first. The MAIN was adapted for online administration closely following the offline version of the instrument. In the offline administration of the assessment, the participant chooses one of three envelopes presented and then proceeds to take out the story picture strip from that envelope. This way the investigator supposedly cannot know which story the participant will choose nor can the investigator see the pictures while the participant is describing them. However, in the online version the investigator is required to share their screen with the PowerPoint slides containing the picture sequences, which makes it impossible to maintain the pretense of the investigator not knowing which story was chosen and what the pictures look like. The instructions were thus modified to include the line: "While you are talking, imagine that I cannot see the pictures." Similarly, in the online version, it is the investigator who controls the progression of the picture sequence, unlike in the in-person administration. The initial slide shows three envelopes, from which the participants need to choose. The following slide displays an embedded video with the full story unfolding, one picture at a time. Three slides each containing a picture pair were then shown, mimicking the unfolding of two pictures at the time. For the comprehension part of the test, a slide showing all six pictures, was displayed with a red frame around the pictures relevant to the question being asked.

For the purposes of the current analyses, the accuracy of all six case markings (nominative, genitive, dative, accusative, instrumental and locative) in obligatory contexts was scored and examined only on nouns (case inflection on adjectives and other lexical items were not included in the analysis).

TABLE 5 Case marking raw scores and percentages per condition and story.

	Heritage speakers		Monolinguals	
	Cat	Baby Birds	Cat	Baby Birds
Nominative	99.2% (2/279)	97.5% (8/320)	100% (212)	99.4 (1/196)
Genitive	98.9% (1/94)	100% (66)	100% (69)	98.6 (1/72)
Dative	100% (9)	92.8% (1/14)	100% (11)	100% (9)
Accusative	93.8% (19/309)	96.4% (7/199)	99.5% (1/248)	100% (168)
Instrumental	93.4% (3/46)	96.6% (1/30)	100% (31)	100% (23%)
Locative	97.4% (2/79)	97% (2/70)	100% (72)	100% (66)
Total score	96.7% (27/814)	97.3% (19/700)*	99.8% (1/643)	99.6% (2/534)
P-value	0.496		0.459	

*Total score including a single correct instance of the vocative case.

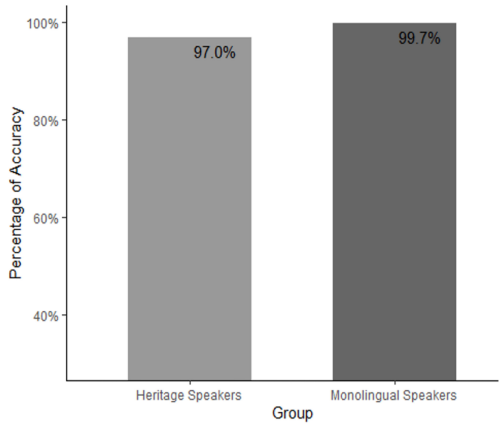

FIGURE 1
Percentage of case marking accuracy per group.

An item for analysis equaled a case inflection on a noun. Each token rather than type of noun case inflection produced within a participant's narrative was considered an item. Items were scored 1 for accurate – on target inflection, or 0 for inaccurate – non-target inflection (encompassing omission, substitution and novel marking). The inflection accuracy per case was also analyzed.

Results

As there was no significant difference in the scores between the stories, the case accuracy scores were merged for analysis (Table 5).

Motivated by the first research question, a logistic mixed effect model was fitted using the lme4 package in R (Bates et al., 2015). This model predicted accuracy based on the group (monolingual or heritage speaker) while allowing varying intercepts for subjects and items. The results from this model are found in Table 6. Group was shown to be a significant predictor of accuracy, $\beta = 2.66$ ($SE = 0.72$), $z = 3.67$, $p = 0.0001$.

Looking at the accuracy rates between the two groups based on raw data, it is possible to deduce the direction of this difference (see Figure 1). Monolingual speakers had an almost perfect performance (99.7%, 3/1174), while the heritage speakers had a slightly lower accuracy rate (97.0%, 46/1465). Using the emmeans package in R, the odds ratio of the two groups was calculated (OR = 0.0698, 95% CI:0.016,0.289) (Lenth et al., 2021). With the help of the R package effect size and applying the "Chen, 2010" rule, the effect size was estimated to be very small (Chen et al., 2010; Ben-Shachar et al., 2020).

The second research question focused on the heritage speaker group. In order to investigate whether accuracy in the case marking of heritage speakers can be predicted by their usage of Bosnian, quantified through the Bosnian Usage Score (BUS), another logistic mixed effect model with random intercepts for participants was fitted (see Table 7). AoO and SES were added to the model as additional predictors. The model showed that the only significant predictor of accuracy was the BUS, with a higher score increasing the probability of higher case marking accuracy, $\beta = 0.30$ ($SE = 0.09$), $z = 3.20$, $p \leq 0.001$.

Case marking error analysis

Both omission and substitution errors were observed in the case marking of heritage speakers. There was one instance of novel marking – the case suffix did not correspond to any existing case marking suffixes. In some cases of omission, it was impossible to tell whether the error was a genuine case marking omission or a substitution with the nominative case, which takes a null ending. In other cases, this differentiation was possible: (1) when the NP contained other elements such

TABLE 6 Regression coefficients of group on case marking accuracy.

	Fixed effects				Random effects	
					By participant	By item
Parameters	**Estimate**	***SE***	**Wald *z***	***p***	***SD***	***SD***
(Intercept)	3.82	0.36	10.53	<0.001	1.10	0.16
Group	2.66	0.72	3.67	<0.001		
Observations	2691					
Model equation	Accuracy ~ Group + (1 | ID) + (1 | Trial)					

as adjectives, demonstratives, etc. which exhibited congruency and were overtly inflected in the nominative; (2) with nouns belonging to paradigms in which the nominative form is overtly marked. However, having no means of verifying the substitution claim in other instances, all forms with zero marking in contexts requiring overtly inflected, non-nominative case, were considered omissions (example 11).

(11) Mačka skače na leptir-Ø*
Cat$_{NOM.F.SG}$ jump$_{3SG.PRS}$ on butterfly- Ø
"The cat is jumping on the butterfly"

Within the substitution errors, there emerged two specific subgroups of errors. The first subgroup occurred when the case assigners were the so-called "two-way" prepositions. These prepositions can assign either accusative or locative/instrumental case, depending on whether they express directionality (accusative) or location (locative/instrumental). Our heritage speakers were found to misselect the appropriate case marking, using locative instead of accusative and vice versa (examples 12 and 13 respectively).

(12) . . .Pala njegova lopt-a
fell3PSG.PST.PERF his ball-NOM.SG.F
u jezer-u*
in lake-LOC.SG.N
. . ."His ball fell in the lake"

(13) . . .Vidi rib-e
see3PSG.PRS.PERF fish-ACC.SG.F
u kant-u*
in bucket-ACC.SG.F
. . ."Sees the fish in the bucket"

The second subgroup involved the substitution of the accusative case with the nominative case and vice-versa. The former type of substitution was more prevalent than the latter (9/15). In these instances, speakers assigned the nominative case to a direct object, and as noted above, this was apparent through the case agreement of the other NP elements or the overtly marked nominative form of the noun (example 14).

(14) . . .Jedn-a mac-a
one- NOM.SG.F cat-NOM.SG.F
koj-a je
who-F beAUX.PRS.3SG
ugleda-la lijep-i
see-PST.PTCP.F beautiful- NOM.SG.MASC
žut-i leptir-Ø*
yellow- NOM.SG.MASC butterfly- NOM.SG.MASC
. . ."A cat who saw a beautiful yellow butterfly"

Discussion

In the first study to focus on case marking in heritage speakers of a lesser-studied language, Bosnian, our HL participants showed an exceptionally high overall case marking accuracy, at 97% correct. This result is in line with previous findings (Bolonyai, 2002; Hlavac, 2003; Rothweiler et al., 2010; Schmitt, 2010). Our monolingual participants, Bosnian speakers from Bosnia, performed at ceiling. Group was a significant predictor of case marking accuracy, therefore our initial prediction of lower accuracy of case marking in heritage speakers was confirmed. However, the effect size was small, so caution should be exercised when interpreting the magnitude of this difference. The second prediction concerned the nature of case marking in heritage languages, namely the restructuring of the heritage case system through omission of overt case markings and reduction of case oppositions. Both these types of phenomena were observed in the current data.

Error types

Nominative-accusative substitution

In line with previous studies, our heritage speakers produced omission and substitution errors (Bolonyai, 2002; Polinsky, 2006; Rothweiler et al., 2010; Gagarina, 2011; Montrul, 2015). The omission error occurred mostly with the direct object, which was supposed to carry an overt accusative marking. In all instances of omission of the direct object case, the null marked form corresponded to

TABLE 7 Regression coefficients of Bosnian Usage Score (BUS) on case marking accuracy of heritage speaker.

	Fixed effects				Random effects
					By participant
Parameters	**Estimate**	***SE***	**Wald *z***	***p***	***SD***
(Intercept)	−2.96	1.98	− 1.49	0.13	0.73
BUS	0.30	0.09	3.20	<0.001	
Years of Education	0.13	0.09	1.33	0.18	
Age of Onset	−0.10	0.16	−0.63	0.52	
Observations	1514				
Model equation	Accuracy ~ BUS + Years of Education + Age of Onset + (1 | ID)				

10.3389/fpsyg.2022.832831

the nominative form. When categorizing these errors, the more conservative estimate that these were omission errors was used, however, one could also argue that these were instances of nominative-accusative substitution, at least for the first declension feminine and accusative nouns. Such an interpretation perhaps holds some merit. In all cases of overt substitution of accusative with nominative, it was the direct object which erroneously took the nominative marking. This could be interpreted as a trend toward the leveling of nominative and accusative cases, where direct objects are assigned the nominative case, similarly to heritage Russian. This case system reduction could be motivated by the overwhelming presence of nominative-accusative syncretism in some BCMS noun declension paradigms. While the spreading of syncretism has been argued to underlie the changes in case marking observed in Heritage Slavic speakers by Łyskawa and Nagy (2020), the number of errors produced by our participants is too small and does not cover nouns from different declensions for us to make any firm conclusions.[7]

Two-way preposition case assignment

As previously observed in adult and child heritage Russian speakers and adult heritage BCMS speakers, Bosnian heritage speakers in the current study also exhibited some difficulty with two way prepositions and accusative and locative alternation (for BCMS: Hansen et al., 2013; Hansen, 2018; for Russian: Isurin and Ivanova-Sullivan, 2008; Schwartz and Minkov, 2014). The accusative and locative were used interchangeably and indiscriminately with the two-way prepositions. This indicates a disregard for the accusative-locative distinction maintained by whether the preposition assigns the meaning of direction or location respectively. The two-way preposition expressing the same distinction with an accusative-dative alternation also exists in German, suggesting that this lack of discrimination is not influenced by the dominant language. Isurin and Ivanova-Sullivan (2008) suggest that such errors emerge due to the "reanalysis of case functions such as direction, location, means" (p. 81). There does seem to be some consistency in the way these errors were made by our participants, with substitutions involving two-way prepositions rarely implicating those cases not assigned by the preposition.

One interpretation may be that there exists a productive rule in the heritage grammar in which two-way prepositions assign their respective cases, but the distinction between the meaning of the cases is unclear to the speakers, due to reanalysis. Tentatively, it could be suggested that the relative lateness in locative emergence as documented for monolingual Croatian children and bilingual Russian-German children does not allow for a long enough rehearsal period (Kovačević et al., 2009; Gagarina, 2011). Note that this is one of the least frequently used cases in BCMS, both in adult input and child usage (Lukatela et al., 1980; Kovačević et al., 2009), and prone to errors in children as old as 6 (Hržica and Lice, 2013; Hržica and Peretić, 2015). Looking into the interaction of lateness in the locative emergence and the amount of input available for heritage speakers to form rules regarding two-way preposition usage could be the first step in solving this puzzle (Schwartz and Minkov, 2014). This suggestion is also brought forth by Klinge (2010) who observed a higher frequency of "divergent uses" (p. 144) for German two-way prepositions compared to one-way prepositions by German-French bilingual children. As Polinsky (2006) suggested, another possibility could be that heritage speakers retain "chunks" without having a productive rule in place which would allow for generalization. In this case a chunk would consist of a preposition and a noun (either in accusative or locative) which has been memorized from the input and is utilized at random without a productive rule which would determine the correct selection of the case based on the direction/location distinction. Future studies could further investigate this explanation *via* more constrained tasks.

Within noun phrase case mismatch

Instances of case mismatch within the NP were observed in our sample of Bosnian heritage speakers, in line with findings from both Hlavac (2003), Hansen et al. (2013) and Hansen (2018). In the example (15), we can see an instance of both case and number mismatch occurring within a single NP. This could potentially be construed as a transfer from German in which case marking is overt on articles, adjectives and pronouns, while the nouns remain largely unmarked (Blake, 2001).

(15)	. . .Jedna	ptica	sa
	$\text{One}_{\text{NOM.F.SG.}}$	$\text{bird}_{\text{NOM.F.SG.}}$	with
	svojom	bebe*	
	$\text{her.own}_{\text{INS.F.SG}}$	$\text{baby}_{\text{NOM.F.PL}}$	
	. . ."A bird with her babies"		

Unintegrated noun phrase heads

Hlavac (2003) noted instances of both English-origin nouns which were integrated into a BCMS clause by taking on overt markings of BCMS (e.g., u hospital-u) as well as English-origin nouns which were unintegrated and simply inserted into the BCMS clause. In the present data only unintegrated German nouns were found. These unintegrated head nouns had a similar effect on the other congruent NP elements to the one observed by Hlavac: the quantifier in (16) and the deictic pronoun in (17) are both used

7 As suggested to us by Boban Arsenijević, it is also possible that animacy could play some role here: the difference between accusative and nominative forms is observed only for animate, but not inanimate nouns belonging to class 1 declension. If errors are more frequent in this class than in others, such pattern can be interpreted as due to the lack of representation of the differential object marking triggered by animacy. We hope to explore this line of interpretation in future research, when more data is available.

10.3389/fpsyg.2022.832831

with an unmarked case. Since the unintegrated head noun that governs the other NP elements lacks the appropriate language specific features such as gender and case, the agreement is unable to be checked and the dependents appear in unmarked forms.

(16) . . .Na prvoj slici ima
On first$_{LOC.F.SG.}$ picture$_{LOC.F.SG}$ have$_{3SG}$
jedan baum
one$_{NOM.M.SG}$ baum
. . ."On the first picture there is a tree (baum)"

(17) I ona hoće da
And she want$_{3SG.PRS}$ COMP
ganja to schmetterling*
chase$_{INF}$ that$_{NOM.N.SG.}$ schmetterling
"And she wants to chase that butterfly (schmetterling)"

This pattern is far from unique to heritage BCMS: see e.g., Putnam et al. (2021) for discussion of similar examples from other HLs, and possible explanations.

Other errors related to case

Patterns observed in some of our participants can be connected to case morphology in a more implicit manner. The example (18) shows an error in gender assignment: the speaker mistakenly assigns the feminine gender to the neuter noun *gnijezdo* "nest." There are two indicators of this error: (1) the reflexive possessive preceding the noun is feminine; (2) the incorrect locative form of the noun corresponds to the locative form found in the paradigm of feminine nouns ending in -a. These two mistakes jointly suggest that the speaker believes that citation form of this noun is something similar to *gnijezda**. She still maintains the gender agreement within the NP and assigns the correct case and number to the noun, but due to the erroneous gender assignment, an incompatible declension paradigm is applied resulting in a distinctly invalid noun form. Case and gender are related categories, thus errors in gender marking might reveal deeper understanding of the processes of heritage changes in the nominal morphology, in general, and case, in particular.

(18) . . .sjedile su
. . .sit$_{3PL.PTCP}$ be$_{AUX\cdot 3PL}$
male ptičice
little$_{NOM.F.PL}$ bird$_{NOM.F.PL}$
u svojoj gnijezdi*
in their-own$_{LOC.SG.F}$ nest$_{LOC.SG.F}$
"little birds were sitting in their nest"

General discussion

As evident in the examples above, there is a systematicity to the errors observed in the data across our participants. These could indicate a systematic restructuring of the underlying heritage grammar. Patterns of case leveling and substitutions similar to the ones found here have been observed in heritage speakers of the related heritage Russian (Polinsky, 2006, 2008; Isurin and Ivanova-Sullivan, 2008). These cross-linguistic findings lend credibility to the argument of heritage grammar restructuring (Putnam et al., 2021). However, we would certainly be remiss not to take into account the possibility that these patterns actually originate as a variation found in the homeland grammars, but are exacerbated by the specific conditions of heritage language acquisition (Flores, 2015; Bousquette and Putnam, 2020; Łyskawa and Nagy, 2020; Wiese et al., 2022). For example, the interchangeable use of accusative and locative with the two way prepositions observed in our data could be associated with the non-standard varieties of BCMS. In some dialects spoken in Serbia and Montenegro, accusative rather than the normative locative/instrumental is consistently used (Ivić, 1985). Thus the locative in "ja sam u trećem razredu" (I am in the third grade) in the standard form of BCMS can vary with the accusative: "ja sam u treći razred," while the instrumental denoting place in the PP "pod slamom" varies with the accusative "pod slamu" (under the hay) (Ivić, 1985, p. 104). While the locative or instrumental denoting place are not used in these dialects, it is possible that speakers of such dialects occasionally use them, especially when in contact with speakers of the standard form, as they do not have the grammatical representation of the locative/instrumental (we thank Boban Arsenijević for this insight). This may result in hypercorrection, with the result of locative (or instrumental) being used in the environments where accusative should be used. Such a pattern has been informally observed by the third author in Albanian-Montenegrin bilinguals living in Montenegro: "Išao sam u Podgorici" (I went to Podgorica). Unfortunately, we are missing key information in order to make any sound judgment of this argument. There is no large corpora of spoken BCMS and their respective dialects to give us an insight into the possible variations currently present in the homeland variants. Our own sample of homeland speakers was fairly homogeneous: almost all of them came from central Bosnia, while our heritage speakers may have been exposed to a more diverse input as they are more likely to be surrounded by speakers from different areas within the BCMS dialect continuum spoken by the diaspora communities from former Yugoslavia. However, our questionnaire did not encompass questions on possible dialectal variations that our participants were in contact with. Such an analysis will hopefully be possible in the future, as there is a great need (both for clinical and research purposes) for a corpus representative of the diversity present in BCMS dialects, especially in the spoken register.

The role of linguistics and non-linguistics factors: Language use, age of L2 onset and socio-economic status

Our measure of language use, the Bosnian Usage Score (BUS), was shown to be a significant predictor of case accuracy. These results are consistent with the previous research, which has demonstrated that HL input and use affect the development and maintenance of its inflectional morphology (Gathercole and Thomas, 2009; Putnam and Sánchez, 2013; Unsworth, 2015, 2016). Neither the AoO of L2 German nor the SES status as conveyed through the number of years of education were significant predictors of Bosnian case marking accuracy.

The lack of an AoO effect is not an isolated result, as AoO has not always been found to predict inflectional morphology accuracy in heritage speakers (Kupisch et al., 2016; Makrodimitris and Schulz, 2021). Our findings regarding case marking are in direct contrast to those of Gagarina and Klassert (2018), which showed AoO to be a significant predictor of case marking in heritage Russian. However, this discrepancy could be explained by a crucial difference between the populations: in contrast to our adult heritage speakers, Gagarina and Klassert's participants included very young children, aged 26–98 months. The acquisition of case morphology in monolingual BCMS speaking children is a long process with errors still being registered at preschool age, at 5–6 years old (Vrsaljko and Paleka, 2018). It is thus possible that long-term continuous input and use exert a greater influence over the adult HL grammar outcomes, outweighing the effect of the L2 AoO. The fact that both studies found use at home (this factor constituting a large portion of the BUS) a significant predictor of case marking accuracy, lends credibility to this proposal. The lack of SES effect on heritage Bosnian case marking accuracy cannot be explained through the homogeneity of the group as in Armon-Lotem et al. (2011). Our group of heritage speakers was fairly heterogeneous in terms of their education, with the years of education ranging from 10 to 19. A more comprehensive SES score including additional SES variables such as the participant's income bracket or the level of parents' education may have constituted a better proxy for the SES as a whole and showed different results. Note also that in her study of the relationship between SES and proficiency scores in bilingual children, De Cat (2020) found that SES advantage only existed in cases of considerable, above-average amount of exposure. The current sample size is too small to perform a separate and reliable analysis for the participants with an above-average BUS and test whether our data would support this finding as well. It is also worth bearing in mind that SES may affect language outcomes in children and adults differently: both Armon-Lotem et al. (2011) and De Cat (2020) focus on children, making their results less applicable to our investigation of HL in adults.

Summary and conclusion

The current study contributes to a growing body of research mapping out the characteristics of heritage BCMS, relying on the online administration of a narrative task. No study so far has examined structured narratives of heritage Bosnian, nor focused specifically on its nominal morphology. Whereas all of the heritage Bosnian studies previously conducted were of a qualitative nature, this study additionally provides quantitative evidence for divergent outcomes in heritage Bosnian grammar. The meticulously controlled tools and methods used to compile the corpus presented here allow for replicability lacking in earlier research. Moreover, prior studies either did not take into account linguistic background (AoO, HL use and input) (Raecke, 2006; Hansen et al., 2013; Hansen, 2018) or, if they did, no attempts were made to relate these factors systematically to the linguistic phenomena observed in the heritage language (Hlavac, 2003). The current study not only gathered relevant data on a range of linguistic factors such as AoO, HL input and usage and non-linguistic factors such as SES across participants, but this information was also utilized to quantitatively evaluate which of these factors are crucial to the development of specific HL domains. From a broader perspective, the results of the study enrich our understanding of the language change and add knowledge on the directions of HL restructuring.

Limitations and future directions

One issue to be considered is the appropriateness of performing a comparison between adult heritage speakers and their monolingual age-matched counterparts in the homeland. According to Polinsky (2018) this comparison might not be particularly useful due to the difference in the input received, especially in adolescence and adulthood. The monolingual environment allows for the development of language novelties that may be unattainable for the heritage speaker group due to the difference in the amount and quality of input. This argument is legitimate, and perhaps an even more informative profile of adult Bosnian heritage speakers could be drawn up by simultaneously investigating the language of monolingual Bosnian children, child heritage speakers or first-generation immigrants. Examining the development of Bosnian in both monolingual Bosnian children and their heritage speaker counterparts of various age groups could provide a clearer picture of the trajectory of language acquisition which leads to the outcomes witnessed in the current data. The same applies for the language of first-generation immigrants, whose input helped shape the language of the heritage speakers. However, limited resources prevented these avenues of research from being pursued here. Yet we believe that the data presented here will be helpful in advancing the field of heritage language research for under-researched languages, such as Bosnian.

Lastly, in the current sample, only the participants from Austria reported attending Bosnian classes either in school or at the University. Austria has a well-documented availability of the so-called "mother language classes" (*Muttersprachlicher Unterricht)* at both the lower and upper levels of secondary schooling (Carnevale et al., 2007). The access to minority language instruction is not as widely available in public schools in Germany. This is apparent through anecdotal evidence reported in the media, as well as official research performed by the information platform Mediendienst Integration (Dribbusch, 2020; Mediendienst Integration, 2020; Voßkühler, 2021). BCMS is offered as a language course in public schools in only 5 out of 16 German states (Hamburg, Hessen, Rheinland-Pfalz, Sachsen-Anhalt, and Schleswig-Holstein). The lack of BCMS minority language support by German public schools is also evident in the information provided by the German participants in this study. The research into the effects of HL input and use on the development of the HL is therefore valuable, as it can ultimately help influence governmental policies on the availability of education in heritage languages. This greatly determines access to HL input during the crucial school years of child heritage speakers.

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 Deutsche Gesellschaft für Sprachwissenschaft (DGfS). The participants provided their written informed consent to participate in this study.

Author contributions

IJ, NG, and AP designed the study. IJ collected and analyzed the data. NG and her colleagues developed the narrative assessment, which was published in 2019. IJ and AP translated and adapted the narrative assessment into Bosnian. NG and AP wrote the article together with IJ. All authors contributed to the article and approved the submitted version.

Acknowledgments

We thank our participants in Germany, Austria, and Bosnia for taking part, Katrin Bente Karl and her team for providing us with the online version of MAIN, Anika Lloyd-Smith and Tanja Kupisch for giving their consent for the adaptation of their Language Use Score questionnaire. We also thank Katarina Bujandric and Anna Artemova for advice on the statistical analyses, and Branko Stanković and Boban Arsenijević for valuable feedback.

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

References

Alexander, R. (2006). *Bosnian, Croatian, Serbian, a Grammar: With sociolinguistic commentary*. Madison: University of Wisconsin press.

Anderson, R. T. (1999). Loss of gender agreement in L1 attrition: preliminary results. *Biling. Res. J.* 23, 389–408. doi: 10.1080/15235882.1999.10162742

Armon-Lotem, S., Walters, J., and Gagarina, N. (2011). The impact of internal and external factors on linguistic performance in the home language and in L2 among Russian-Hebrew and Russian-German preschool children. *Linguist. Approach. Biling.* 1, 291–317. doi: 10.1075/lab.1.3.04arm

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Statist. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Ben-Shachar, M. S., Lüdecke, D., and Makowski, D. (2020). Effectsize: estimation of effect size indices and standardized parameters. *J. Open Source Softw.* 5:2815. doi: 10.21105/joss.02815

Blake, B. J. (2001). *Case*, 2nd Edn. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139164894.006

Bolonyai, A. (2002). Case systems in contact: syntactic and lexical case in bilingual child language. *Southwest J. Linguist.* 21, 1–36.

Bousquette, J., and Putnam, M. T. (2020). Redefining language death: evidence from moribund grammars. *Lang. Learn.* 1, 188–225. doi: 10.1111/lang.12362

Carnevale, C., de Cillia, R., Krumm, H.-J., and Schlocker, E. (eds). (2007). *Language education policy profile: Country report Austria*. Vienna: Austrian Federal Ministry for Education.

Ćatibušić, B. (2019). Minority language development in early childhood: a study of siblings acquiring Bosnian and English in Ireland. *TEANGA* 10, 139–161. doi: 10.35903/teanga.v10i0.75

Chen, H., Cohen, P., and Chen, S. (2010). How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies. *Commun. Statist. Simulat. Computat.* 39, 860–864.

Cobo-Lewis, A. B., Eilers, R. E., Pearson, B. Z., and Umbel, V. C. (2002). "Chapter 6: interdependence of Spanish and English knowledge in language and literacy among bilingual children," in *Language and literacy in bilingual children*, eds K. Oller, and R. Eilers (Bristol: Multilingual Matters), 118–132. doi: 10.21832/9781853595721-007

Czapka, S., Topaj, N., and Gagarina, N. (2021). A four-year longitudinal comparative study on the lexicon development of Russian and Turkish heritage speakers in Germany. *Languages* 6:27. doi: 10.3390/languages6010027

De Cat, C. (2020). Predicting language proficiency in bilingual children. *Stud. Second Lang. Acquisit.* 42, 279–325. doi: 10.1017/S0272263119000597

Dribbusch, B. (2020). *Deutsch als Zweitsprache: Kein Grund zur Panik*. Berlin: Die Tageszeitung.

Embick, D., White, Y., and Tamminga, M. (2020). Heritage languages and variation: identifying shared factors. *Bilingualism* 23, 21–22. doi: 10.1017/S1366728919000476

Flege, J. E., Yeni-Komshian, G. H., and Liu, S. (1999). Age constraints on second-language acquisition. *J. Mem. Lang.* 41, 78–104. doi: 10.1006/jmla.1999.2638

Flores, C. M. M. (2015). Understanding heritage language acquisition. Some contributions from the research on heritage speakers of European Portuguese. *Lingua* 164, 251–265.

Gagarina, N. (2011). "Acquisition and loss of L1 in a Russian-German bilingual child: a case study," in *Monolingual and Bilingual Path to Language*, ed. S. Cejtlin (Moskau: Jazyki slavjanskoj kul_x0019_ tury), 137–163.

Gagarina, N. (2017). "Monolingualer und bilingualer Erstspracherwerb des Russischen: ein Überblick," in *Handbuch des Russischen in Deutschland: Migration – Mehrsprachigkeit – Spracherwerb*, eds N. Wulff and K. Witzlack-Makarevich (Berlin: Frank & Timme), 393–410.

Gagarina, N., and Klassert, A. (2018). Input dominance and development of home language in Russian-German bilinguals. *Front. Commun.* 3:40. doi: 10.3389/fcomm.2018.00040

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., et al. (2012). MAIN: multilingual assessment instrument for narratives. *ZAS Papers Linguist.* 56:155.

Gagarina, N., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Bohnacker, U., et al. (2019b). MAIN: multilingual assessment instrument for narratives – revised. *ZAS Papers Linguist.* 63:20. doi: 10.21248/zaspil.63.2019.516

Gagarina, N., Bohnacker, U., and Lindgren, J. (2019a). Macrostructural organization of adults' oral narrative texts. *ZAS Papers Linguist.* 62, 190–208. doi: 10.21248/zaspil.62.2019.449

Gamlen, A. (2019). "Intercultural borders in europe and its emulators," in *Human Geopolitics: States, Emigrants, and the Rise of Diaspora Institutions*, (Oxford: Oxford University Press), 125–158.

Gathercole, V. C. M., and Thomas, E. M. (2009). Bilingual first-language development: dominant language takeover, threatened minority language take-up. *Bilingualism* 12, 213–237. doi: 10.1017/S1366728909004015

Hansen, B. (2018). "On the permeability of grammars: syntactic pattern replications in heritage Croatian and heritage Serbian spoken in Germany," in *Diachronic Slavonic Syntax: The Interplay between Internal Development, Language Contact and Metalinguistic Factors*, eds B. Hansen, J. Grković-Major, and B. Sonnenhauser (Berlin: De Gruyter Mouton), 125–160. doi: 10.1515/9783110531435-006/html

Hansen, B., Romić, D., and Kolaković, Z. (2013). Okviri za istraživanje sintaktičkih struktura govornika druge generacije bosanskoga, hrvatskoga i srpskoga jezika u Njemačkoj. *Izvorni Znanstveni Članak* 15, 9–45.

Hart, B., and Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MA: Paul H Brookes Publishing.

Hartmann, J. M., and Milicevic, N. (2009). "Case alternations in serbian existential," in *Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Information Structure*, eds G. Zybatow, U. Junghanns, D. Lenertova, and P. Biskup (Frankfurt: Peter Lang), 131–142.

Hlavac, J. (2003). *Second-Generation Speech: Lexicon, Code-Switching and Morpho-Syntax of Croatian-English Bilinguals*. Bern: Peter Lang.

Hoff, E. (2006). How social contexts support and shape language development. *Dev. Rev.* 26, 55–88. doi: 10.1016/j.dr.2005.11.002

Hržica, G., and Lice, K. (2013). Morfološke pogreške u uzorcima govornog jezika djece urednog jezičnog razvoja i djece s posebnim jezičnim teškoćama. *Hrvatska Rev. rehabil. Istraživanja* 49, 65–77.

Hržica, G., and Peretić, M. (2015). "Što je jezik?," in *Priručnik za Prepoznavanje i Obrazovanje Djece s Jezičnim Teškoćama*, ed. J. K. Kraljevic (Zagreb: Edukacijsko-rehabilitacijski fakultet Sveučilišta u Zagrebu).

Isurin, L., and Ivanova-Sullivan, T. (2008). Lost in between: the case of Russian heritage speakers. *Heritage Lang. J.* 72–104. doi: 10.46538/hlj.6.4

Ivić, P. (1985). *Dijalektologija Srpskohrvatskog Jezika - Uvod U Štokavsko Narečje*, 2nd Edn. Matica srpska: Novi Sad.

Klinge, S. (2010). "Chapter 6. Prepositions in bilingual language acquisition," in *Two first languages*, ed. J. Meisel (Berlin: De Gruyter Mouton), 123–156. doi: 10.1515/9783110846065.123

Kovačević, M., Palmović, M., and Hržica, G. (2009). "The acquisition of case, number, and gender in Croatian," in *Development of Nominal Inflection in First Language Acquisition: A Cross-Linguistic Perspective*, eds U. Stephany and M. D. Voeikova (Berlin: De Gruyter Mouton), 153–178. doi: 10.1515/9783110217117.153/html

Kupisch, T. (2013). A new term for a better distinction? A view from the higher end of the proficiency scale. *Theor. Linguist.* 39, 203–214. doi: 10.1515/tl-2013-0012

Kupisch, T. (2019). "2L1 Simultaneous bilinguals as heritage speakers," in *The Oxford Handbook of Language Attrition*, eds M. S. Schmid and B. Köpke (Oxford: Oxford University Press).

Kupisch, T., Belikova, A., Özçelik, Ö, Stangen, I., and White, L. (2016). Restrictions on definiteness in the grammars of German-Turkish heritage speakers. *Linguist. Approach. Biling.* 7, 1–38. doi: 10.1075/lab.13031.kup

La Morgia, F. (2015). "Assessing the relationship between input and strength of language development: a study on Italian–English bilingual children," in *Language*, eds C. Silva-Corvalán and J. Treffers-Daller (Cambridge: Cambridge University Press).

Lenth, R. V., Buerkner, P., Herve, M., Love, J., Riebl, H., and Singmann, H. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means (1.7.0) [Computer software]*. Available online at: https://CRAN.R-project.org/package=emmeans\protect\T1\textsection (accessed November 12, 2014).

Lloyd-Smith, A. (2020). *Heritage Bilingualism and the Acquisition of English as a Third Language*. Ph.D Dissertation. Germany: University of Konstanz.

Lohndal, T., Rothman, J., Kupisch, T., and Westergaard, M. (2019). Heritage language acquisition: what it reveals and why it is important for formal linguistic theories. *Lang. Linguist. Compass* 13:e12357. doi: 10.1111/lnc3.12357

Lukatela, G., Gligorijević, B., Kostić, A., and Turvey, M. T. (1980). Representation of inflected nouns in the internal lexicon. *Memory Cognit.* 8, 415–423. doi: 10.3758/BF03211138

Łyskawa, P., and Nagy, N. (2020). Case marking variation in heritage Slavic languages in Toronto: not so different. *Lang. Learn.* 70, 122–156.

Makrodimitris, C., and Schulz, P. (2021). Does timing in acquisition modulate heritage children's language abilities? Evidence from the Greek LITMUS Sentence Repetition Task. *Languages* 6:49. doi: 10.3390/languages6010049

Mediendienst Integration. (2020). *Wie Verbreitet ist Herkunftssprachlicher Unterricht?*. Available online at: https://mediendienst-integration.de/artikel/wie-verbreitet-ist-herkunftssprachlicher-unterricht.html#:~:text=Sieben%20Bundesl%C3%A4nder%20konnten%20Angaben%20zur,mehr%20als%20im%20Schuljahr%20zuvor (accessed August 6, 2020).

Meisel, J. M. (2009). Second language acquisition in early childhood. *Z. Sprachwissenschaft* 28, 5–34. doi: 10.1515/ZFSW.2009.002

Meisel, J. M. (2011). "Neural maturation and age: opening and closing windows of opportunities," in *First and Second Language Acquisition: Parallels and Differences*, ed. J. M. Meisel (Cambridge: Cambridge University Press), 202–239.

Montrul, S. (2002). Incomplete acquisition and attrition of Spanish tense/aspect distinctions in adult bilinguals. *Bilingualism.* 5, 39–68. doi: 10.1017/S1366728902000135

Montrul, S. (2015). *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press.

Paradis, J. (2011). Individual differences in child English second language acquisition: comparing child-internal and child-external factors child-internal and child-external factors. *Linguist. Approach. Biling.* 1, 213–237.

Paradis, J., Nicoladis, E., Crago, M., and Genesee, F. (2011). Bilingual children's acquisition of the past tense: a usage-based approach. *J. Child Lang.* 38, 554–578. doi: 10.1017/S0305000910000218

Polinsky, M. (2006). Incomplete acquisition: American Russian. *J. Slavic Linguist.* 14, 191–262.

Polinsky, M. (2008). "Heritage language narratives," in *Heritage Language Education: A New Field Emerging*, eds D. M. Brinton, O. Kagan, and S. Bauckus (London: Routledge), 149–164. doi: 10.4324/9781315092997-11

Polinsky, M. (2018). *Heritage Languages and their Speakers*. Cambridge: Cambridge University Press, doi: 10.1017/9781107252349

Polinsky, M., and Scontras, G. (2020). Understanding heritage languages. *Bilingualism* 23, 4–20. doi: 10.1017/S1366728919000245

Putnam, M. T., and Sánchez, L. (2013). What's so incomplete about incomplete acquisition? - A prolegomenon to modeling heritage language grammars. *Linguist. Approach. Biling.* 3, 478–508.

Putnam, M. T., Schwarz, L., and Hoffman, A. D. (2021). "Morphology in heritage languages," in *Cambridge Handbook of Heritage Languages*, eds M. Polinsky and S. Montrul (Cambridge: CUP), 613–643.

Raecke, J. (2006). Hrvatski u Njemačkoj: Njemački s hrvatskim riječima? *Lahor* 2, 151–158.

Randjelovic, I. (2019). *Being and belonging: Stories of Second-Generation Serbian Migrants in Germany and Australia*. Hawthorn, VIC: Swinburne University of Technology.

Rothweiler, M., Chilla, S., and Babur, E. (2010). Specific language impairment in Turkish: evidence from case morphology in Turkish–German successive bilinguals. *Clin. Linguist. Phonet.* 24, 540–555. doi: 10.3109/02699200903545328

Sánchez, R. (1983). *Chicano Discourse: Socio-historic Perspectives*. NewYork, NY: Newbury House Publishers.

Savić, S. (1989). "Dokle smo došli?," in *Interkulturalizam kao Oblik Obrazovanja Dece Migranata Van Domovine: Zbornik radova*, ed. S. Savić (Novi Sad: Institut za južnoslovenske jezike).

Schlund, K. (2006). Sprachliche Determinanten bilingualer Identitätskonstruktion am Beispiel von Deutsch-Jugoslawien der zweiten Generation. *Z. Slawistik* 51, 74–93. doi: 10.1524/slaw.2006.51.1.74

Schmitt, E. (2010). When boundaries are crossed: evaluating language attrition data from two perspectives*. *Bilingualism* 13, 63–72. doi: 10.1017/S1366728909990381

Schwartz, M., and Minkov, M. (2014). Russian case system acquisition among Russian–Hebrew speaking children. *J. Slavic Linguist.* 22, 51–92.

Scontras, G., and Putnam, M. T. (2020). Lesser-studied heritage languages: an appeal to the dyad. *Heritage Lang. J.* 17, 152–155.

Thomas, E., and Gathercole, V. (2005). "Minority language survival: input factors influencing the acquisition of Welsh," in *Proceedings of the 4th International Symposium on Bilingualism*, eds J. Cohen, K. T. McAlister, K. Rolstad, and J. Macswan (Somerville, MA: Cascadilla Press).

Unsworth, S. (2015). "Amount of exposure as a proxy for dominance in bilingual language acquisition," in *Language Dominance in Bilinguals: Issues of Measurement and Operationalization*, eds C. Silva-Corvalán and J. Treffers-Daller (Cambridge: Cambridge University Press), 156–173. doi: 10.1017/CBO9781107375345.008

Unsworth, S. (2016). "Quantity and quality of language input in bilingual language development," in *Bilingualism Across the Lifespan: Factors Moderating Language Proficiency*, eds E. Nicoladis and S. Montanar (Washington, DC: American Psychological Association), 103–122.

Van Osch, B., Hulk, A., Aalberse, S., and Sleeman, P. (2018). Implicit and explicit knowledge of a multiple interface phenomenon: differential task effects in heritage speakers and L2 speakers of Spanish in the Netherlands. *Languages* 3:25. doi: 10.3390/languages3030025

Velupillai, V. (2012). *An Introduction to Linguistic Typology, z.176*. Amsterdam: John Benjamins Publishing Company.

Voßkühler, G. (2021). *Herkunftssprachlicher Unterricht: Nicht mehr nur Deutschstunde*. Berlin: Die Tageszeitung.

Vrsaljko, S., and Paleka, P. (2018). Pregled ranoga govorno-jezičnoga razvoja. *Magistra Iadertina* 13, 139–159.

Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., et al. (2022). Heritage speakers as part of the native language continuum. *Front. Psychol.* 12:717973. doi: 10.3389/fpsyg.2021.717973

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., et al. (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language* 31, 461–479. doi: 10.1177/0142723711409976

frontiers | Frontiers in Communication

TYPE Methods
PUBLISHED 08 March 2023
DOI 10.3389/fcomm.2023.919617

OPEN ACCESS

EDITED BY
Wenchun Yang,
Leibniz Center for General Linguistics (ZAS), Germany

REVIEWED BY
Rachel Ostrand,
IBM Research, United States
Seçkin Arslan,
Centre National de la Recherche Scientifique (CNRS), France

*CORRESPONDENCE
Spyridoula Stamouli
✉ pstam@athenarc.gr

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 13 April 2022
ACCEPTED 15 February 2023
PUBLISHED 08 March 2023

CITATION
Stamouli S, Nerantzini M, Papakyritsis I, Katsamanis A, Chatzoudis G, Dimou A-L, Plitsis M, Katsouros V, Varlokosta S and Terzi A (2023) A web-based application for eliciting narrative discourse from Greek-speaking people with and without language impairments. *Front. Commun.* 8:919617. doi: 10.3389/fcomm.2023.919617

A web-based application for eliciting narrative discourse from Greek-speaking people with and without language impairments

Spyridoula Stamouli[1]*, Michaela Nerantzini[2], Ioannis Papakyritsis[2], Athanasios Katsamanis[1], Gerasimos Chatzoudis[1], Athanasia-Lida Dimou[1], Manos Plitsis[1,3], Vassilis Katsouros[1], Spyridoula Varlokosta[1,4] and Arhonto Terzi[2]

[1]Institute for Language and Speech Processing, Athena Research Center, Athens, Greece, [2]Department of Speech and Language Therapy, University of Patras, Patras, Greece, [3]Department of Informatics and Telecommunications, University of Athens, Athens, Greece, [4]Faculty of Philology, Linguistics Department, National and Kapodistrian University of Athens, Athens, Greece

In this paper we present a web-based data collection method designed to elicit narrative discourse from adults with and without language impairments, both in an in-person set up and remotely. We describe the design, methodological considerations and technical requirements regarding the application development, the elicitation tasks, materials and guidelines, as well as the implementation of the assessment procedure. To investigate the efficacy of remote elicitation of narrative discourse with the use of the technology-enhanced method presented here, a pilot study was conducted, aiming to compare narratives elicited remotely to narratives collected in an in-person elicitation mode from ten unimpaired adults, using a within-participants research design. In the remote elicitation setting, each participant performed the tasks of a narrative elicitation protocol *via* the web application in their own environment, with the assistance of an investigator in the context of a virtual meeting (video conferencing). In the in-person elicitation setting, the participant was in the same environment with the investigator, who administered the tasks using the web application. Data were manually transcribed, and transcripts were processed with Natural Language Processing (NLP) tools. Linguistic features representing key measures of spoken narrative discourse were automatically calculated: linguistic productivity, content richness, fluency, syntactic complexity at clausal and inter-clausal level, lexical diversity, and verbal output. The results show that spoken narratives produced by the same individuals in the two different experimental settings do not present significant differences regarding the linguistic variables analyzed, in sixty six out of seventy statistical tests. These results indicate that the presented web-based application is a feasible method for the remote collection of spoken narrative discourse from adults without language impairments in the context of online assessment.

KEYWORDS

narrative discourse, elicitation, web application, remote assessment, aphasia

1. Introduction

Acquired speech and language disorders are increasingly relevant for a significant percentage of the adult population, given the current aging rate, and they have a direct and severe effect on their quality of life, since they limit daily communication. The effective support of adults with language impairments requires individualized, systematic, and regular intervention by speech and language therapists (SLTs). Timely assessment is an essential step for the identification of their communication abilities and deficits, for prognosis of functional recovery, as well as for the design of individualized intervention plans.

Direct, face-to-face (FTF) clinical services have been considered the gold standard of behavioral appraisal and intervention in the field of speech and language pathology. Effective service delivery requires clinicians to be able to offer real-time directions and feedback cues that are directly responsive to the patient's actions, utterances, or other type of behaviors during the clinical session. However, it could be argued that for some patients with communication disorders, FTF service delivery is not an ideal or a viable option. Especially for patients with significant physical and communication disabilities, FTF sessions may often require disproportionately high levels of physical, cognitive, and emotional effort on the part of the patient, as well as need for caregiver assistance, transport, and increased financial cost.

Teleassessment and telerehabilitation practices have been considered an effective alternative to in-person clinical services, long before the COVID-19 pandemic made them an urgent necessity. A survey of the American Speech-Language-Hearing Association, completed by 476 SLTs, indicated that 64% of the clinicians endorsed providing services *via* telepractice, 37.6% used telepractice for screenings, 60.7% used telepractice for assessment, and 96.4% used it for intervention (American Speech-Language-Hearing Association, 2016). Remote clinical services are particularly relevant in the context of stroke-induced speech and language impairments, such as aphasia, given the high levels of unmet needs and the increased service demands. Additionally, effectiveness of treatment in aphasia is linked to early appraisal and thus it is beneficial for clinicians to have a means to carry out comprehensive assessments of different aspects of the patient's communication abilities, including narrative discourse, with minimal effort from the part of the patient and the clinician, and without the need for transport away from the patient's residence.

To address these needs, a research project was set up, aiming to develop a technology-enhanced platform offering adults with acquired speech and language disorders the opportunity for remote long-term speech and language therapy in their own environment, without the physical presence of an SLT. With the aim to assist STLs in the process of patients' assessment and monitoring of the intervention outcome, the platform integrates a web-based application for the collection of spoken narratives from individuals with speech and language impairments and unimpaired controls. Data collected with this application serve as an evaluation corpus, against which a machine learning system for the automatic assessment of the severity of impairment will be tested for its accuracy and robustness. The system focuses on aphasia, as one of the most complex types of chronic acquired language disorders, affecting the communicative abilities of a significant percentage of the adult population in multiple language modules and modalities.

The purpose of this paper is to present the design, methodological considerations and requirements of the web-based application for the elicitation of spoken narrative discourse from Greek-speaking people with aphasia (PWA) and unimpaired adults. The application allows the online administration of a comprehensive protocol of seven narrative tasks and is designed for remote as well as for in-person administration. Subsequently, we present the first phase of the evaluation of the presented method regarding its efficacy and feasibility in collecting data remotely, as compared to the traditional in-person setting. More specifically, a pilot study which involves only neurotypical adults will be presented, aiming to examine whether the linguistic properties of spoken narratives collected remotely are comparable to the ones collected in a FTF set up using the presented web-based application. The second phase of the method's evaluation will include language impaired participants, namely PWA.

1.1. The study of discourse in aphasia—Narrative discourse

The term *discourse* is commonly used to describe the way in which language is used and structured beyond the level of the sentence to convey an understandable message (Armstrong, 2000; Wright, 2011). The need to collect and study extensive discourse samples from PWA was identified in the late 1970s, when a discrepancy between language performance on standardized aphasia tests and real-life language use for social interactions, termed as *functional communication*, was identified (Holland, 1979).

Aphasia is an acquired language disorder, as a result of focal damage to the left cerebral hemisphere, caused by a cerebral vascular accident (CVA) or a traumatic brain injury (TBI) (Obler and Gjerlow, 1999).[1] Aphasia can affect the production and comprehension of both spoken and written language, at all language levels (phonological, morphological, syntactic and semantic) and to varying degrees, depending on the area and severity of the brain injury (Harley, 2001), causing mild, moderate or severe language impairment.

The aim of aphasia rehabilitation is to improve PWA's functional communication, i.e., individuals' language skills to achieve communication goals in the context of everyday social interactions. Thus, since the late 1970s there is an increasing focus on the study of contextualized language use, since there is an agreement in the aphasia literature that the controlled administration conditions of standardized aphasia assessment protocols mainly focus on isolated components of language - phonology, morphology, syntax and semantics- at word and sentence level, that do not simulate the cognitive requirements and conditions of real-life communication (Holland, 1982; Armstrong,

1 Moreover, another type of aphasia, Primary Progressive Aphasia (PPA), has been identified (Mesulam, 2001). PPA is a neurodegenerative clinical condition associated with frontotemporal dementia (FTD), which primarily affects language functions and is characterized by their gradual loss.

2000; Beeke et al., 2011; Olness and Ulatowska, 2011; Doedens and Meteyard, 2022). In this context, the study of discourse provides a naturalistic and ecologically valid framework for language assessment, which can reveal different aspects of language abilities, as well as weaknesses, of PWA in more natural communicative contexts, unravel interactions between individual language components and assess intervention effects in connected speech (Dietz and Boyle, 2018).

According to the systematic literature review of Bryant et al. (2016), which covers 40 years of the study of discourse in aphasia (1976–2015), studies using discourse analysis methods doubled in the second half of the 1990s, with the most significant increase observed from the late 2000s to 2015. The growing interest in the study of aphasia at the level of discourse over the last two decades has also been influenced by international frameworks and procedures in terms of assessment and rehabilitation, such as the International Classification of Functioning, Disability and Health (ICF) of the World Health Organization (World Health Organization, 2001). In this context, new approaches to the assessment and rehabilitation of communication skills of PWA have begun to develop, focusing on functional communication of individuals and their active participation in daily life, examining not only linguistic factors, such as the nature of the language impairment, but also social and psychological ones, such as social participation, social identity, self-esteem, and mental resilience. Using tools such as systematic observation and assessments of PWA's ability to respond effectively to specific communication situations, like maintaining a dialogue with another interlocutor or recounting a story, the functional approach highlights the level of discourse as a field of study in aphasia, emphasizing the role of the communicative setting and the context in discourse production (Armstrong et al., 2011).

For the study of discourse of PWA and unimpaired controls, different discourse types have been analyzed: exposition, procedural, narrative and conversational discourse. Exposition refers to discourse produced to describe or explain a topic, procedural discourse refers to the description of a process (e.g., how to make a peanut butter and jelly sandwich), narrative represents recounting of a fictional or factual story and conversational discourse is the interactive communication between two or more people. Among these discourse types, narrative is the one that has been more extensively investigated (Bryant et al., 2016). The study of narrative discourse is compatible with the functional approach to the rehabilitation of language disorders in PWA, since narrative is favored as an integral part of human communication, representative of everyday language use. Moreover, narrative offers a controlled framework for the analysis of coherence and discourse organization, provided by global story macrostructure.

1.2. Methods for eliciting narrative discourse from PWA and neurotypical controls

In recent surveys on the use of discourse data for the assessment of aphasia in clinical settings (Bryant et al., 2017; Cruice et al., 2020), as well as in research and clinical settings (Stark et al., 2021b), it has been reported that most clinicians and researchers collect spoken discourse data from PWA and unimpaired controls using a variety of discourse elicitation methods. For instance, free narrative production tasks, such as personal narratives, or structured and semi-structured tasks, such as picture-elicited story production or story retelling, have been frequently used.

The elicitation of personal narratives in the context of the study of aphasia typically involves the narration the person's "stroke story", whereby the participants narrate the events of their stroke. Personal narratives have been employed in aphasia research because of their *multi-functionality*; as personal stories of an individual's experience, they actively involve both functions of narratives, i.e., the referential function, mainly related to the temporal arrangement of events, and the evaluative function, which conveys the narrator's attitudes, emotions, and opinions toward the narrated events (Labov, 1972). The evaluative function of personal narratives is critical for the study of PWA's communicative competence or functionality, since it reflects the intrapersonal and interpersonal function of narration (Olness and Ulatowska, 2011); PWA have the opportunity to talk about themselves and their chronic condition, and, therefore, to establish their sense of identity, self-image and self-expression and, at the same time, to share their experiences with others in the context of a social interaction (Fromm et al., 2011). For these reasons, personal narratives and, especially, illness narratives, are considered a natural context to investigate the degree in which PWA are able to accomplish the intrapersonal and interpersonal functions of storytelling, depending on their language impairments and severity level (Armstrong and Ulatowska, 2007; Olness et al., 2010; Olness and Englebretson, 2011), and, thus, they represent a discourse elicitation task, which is compatible to functional and social models in aphasia assessment and therapy. Moreover, in terms of specific measures of language ability, it has been reported that personal narratives involve more *correct information units* (*CIUs*, Nicholas and Brookshire, 1993) (Doyle et al., 1995), which are considered an objective measure of functional, real-life communication abilities in aphasia (Doedens and Meteyard, 2020), and are characterized by greater complexity, as evaluated by criteria such as vocabulary range, utterance length, subordination, etc., in relation to picture-elicited narratives (Glosser et al., 1988). However, the speech samples include personal events, and are therefore less comparable with each other in terms of informational content, than picture-elicited narratives.

Structured or semi-structured tasks involve two sub-types, which have been widely used in aphasia research: (a) the production of narrative discourse based on visual stimuli and (b) the retelling of stories that have been presented orally. Picture-elicited narratives, termed also as "expositional narratives" (Stark, 2019), or "picture descriptions" (MacWhinney et al., 2011), involve narration based on a single picture or a picture sequence. Tasks of this type do not burden memory, provide a controlled context for narrative production and guide participants to produce comparable stories in terms of narrative macrostructure, lexical elements, main events, and information units. The picture prompts that have been most widely used to elicit narrative discourse include the *Cookie Theft*, which is part of the Boston Diagnostic Aphasia Examination (BDAE) (Goodglass and Kaplan, 1972), the *Picnic*, which is part of the Western Aphasia Battery (WAB-R) (Kertesz, 2006), and

the *Cat Rescue*, the *Birthday Cake*, the *Fight*, and the *Farmer and His Directions* single picture and picture sequences stimuli by Nicholas and Brookshire (1993). It is worth noting that single-picture stimuli, such as the *Cookie theft*, the *Picnic* and the *Cat Rescue*, are not transparently associated with the narrative discourse type. Many studies and aphasia assessment procedures require participants to describe either single pictures or picture sequences as isolated scenes, resulting in the production of a static description of situations and characters, rather than narratives, which typically involve the identification of underlying temporal and causal relations between events (Armstrong, 2000; Wright and Capilouto, 2009). It has been demonstrated that the discourse type elicited (description vs. narrative) as well as the quality of narrative production heavily depend on task instructions given to participants to produce discourse (Olness, 2006; Wright and Capilouto, 2009). Moreover, single pictures have been reported to elicit lower narrative levels, in terms of narrative structure complexity (Lemme et al., 1984; Bottenberg et al., 1985), lower cohesive harmony (Bottenberg et al., 1985) and lower lexical diversity (Fergadiotis and Wright, 2011) than picture sequences.

The second type involves the retelling of a story that participants have previously heard. This type of task does not require speakers to construct the narrative content themselves but requires them to retain the events of the story and their temporal succession, to recall them from memory and reproduce them. The most well-known protocol of this type is the *Story Retell Procedure* (Doyle et al., 2000), which is based on the visual stimuli of Nicholas and Brookshire (1993) and has been evaluated as a reliable and valid method of narrative elicitation (Doyle et al., 2000; McNeil et al., 2001, 2002, 2007). This method, despite the processing demands on working memory posed to participants, produces comparable speech samples, in terms of measures of linguistic performance, to the ones obtained from picture-elicited narratives and procedural discourse tasks (McNeil et al., 2007).

Given the diverse clinical patterns in aphasia and task effects that have been observed in discourse production, regarding parameters such as verbal productivity, fluency, information content, grammatical accuracy, and complexity (Doyle et al., 1998; Stark, 2019), it is recommended (Brookshire and Nicholas, 1994; Armstrong, 2000; Olness, 2006; Stark, 2019; Stark and Fukuyama, 2021) that a combination of elicitation methods should be used to obtain a comprehensive language sample most resemblant of actual language use. To this end, AphasiaBank, the largest repository of multi-lingual and multi-modal data for the study of communication in aphasia, implements a standard protocol for the elicitation of spoken discourse from PWA and unimpaired controls (MacWhinney et al., 2011), which includes personal narratives, picture-elicited narratives, familiar story telling and procedural discourse.[2] Personal narratives include the narration of the PWA's "stroke story" and an "illness story" from neurotypical controls, as well as the story of an important event from both groups. The *Cat Rescue* picture prompt (Nicholas and Brookshire, 1993), as well as the *Refused Umbrella* and the *Broken window* picture sequences are used for the collection of picture-elicited narratives. The AphasiaBank protocol also includes the narration of the traditional fairytale of *Cinderella*, after the revision of a wordless picture book (Grimes, 2005), which is removed prior to narration (Saffran et al., 1989). Finally, PWA are prompted to produce procedural discourse, in which they describe the procedure of making a *Peanut Butter and Jelly Sandwich*.

Moreover, it is evidenced that the amount of data collected per participant is an important issue related to sufficient discourse sampling, so speech samples are representative of participants' language abilities (Brookshire and Nicholas, 1994; Boles and Bombard, 1998; Armstrong, 2000). Brookshire and Nicholas (1994) found that test-retest stability of two measures of spoken discourse, words per minute and percentage of correct information units, increased as sample size increased. They suggested that a speech sample obtained from 4 to 5 different tasks containing a total of 300–400 words per participant represents a sufficient sample size to achieve acceptable high test-retest stability. Boles and Bombard (1998) investigated adequacy of sample size in terms of time duration of the conversational discourse sampled per participant. They found that 10-min samples represented an adequate conversation length to reliably measure the variables of conversational repair, speaking rate and utterance length.

Despite the recommendations for collecting spoken discourse samples with a variety of discourse elicitation methods, it is reported that the number of samples collected for the assessment and analysis of spoken discourse usually ranges from one to four, with most investigators collecting one or two samples per person (Stark et al., 2021a). Several studies highlight significant barriers in implementing discourse data collection methods in clinical practice, the most typical of which is the lack of tools and resources, such as computer software or hardware and audio equipment, as well as inadequate training, knowledge, and skills in discourse collection (Bryant et al., 2017; Cruice et al., 2020; Stark et al., 2021a).

1.3. Remote assessment of language abilities

The elicitation of narratives has been primarily obtained *via* direct FTF set up. However, technological applications for language or communicative skills assessment in educational or clinical settings have some significant advantages over FTF services: (a) they are more practical, since all materials and prompts are integrated in a comprehensive computer environment, and all equipment needed, such as microphone, audio player, and speakers, is built-in and easily accessible *via* the application interface; (b) uniformity of task presentation and administration is facilitated regarding several parameters, such as order of tasks presentation, stimuli presentation order and timing, and instructions format; (c) data storage and management is significantly simplified, since data are stored and logged in the application's backend, allowing the investigators easy data access, tracking and filtering. Moreover, keeping in mind that the ultimate goal of the data collection method reported here is to assess individuals with aphasia, it should be noted that the available research for the delivery of remote clinical services, including teleassessment and telerehabilitation, has produced promising results.

2 The materials and guidelines for the administration of the AphasiaBank protocol can be accessed at https://aphasia.talkbank.org/protocol/english/.

Regarding teleassessment, remote appraisal of language and other communication functions involves either the adaptation of conventional, already available assessment tools for online use, or the development of novel instruments specifically developed for remote administration. In fact, there is a significant body of research on the validity of computerized testing of neurotypical populations (Newton et al., 2013). Although studies conducted in the early 1990s have reported considerably poorer performance on computerized compared to pen and paper tests, the disadvantage found for scores of computerized assessments is now getting smaller and the two procedures are considered comparable (Noyes and Garland, 2008). The discrepancies often reported between the two methods have been attributed to a number of factors, including computer experience, anxiety and participant perceptions toward computerized testing (Newton et al., 2013).

Although there are numerous assessments of speech, language and communication available for aphasia, the literature on the validation of teleassessment versions of these tools is still lacking. However, several studies have demonstrated the potential validity of web-based versions of widely used aphasia assessments. More specifically, Theodoros et al. (2008), Hill et al. (2009), and Palsbo (2007) have compared online vs. FTF administration of the short versions of the BDAE-3 (Goodglass et al., 2001) and the Boston Naming Test (BNT, 2nd edition). Newton et al. (2013) compared computer-delivered and paper-based language tests, including parts of the Comprehensive Aphasia Test (CAT) and the Test of Reception of Grammar (TROG), and Dekhtyar et al. (2020) validated the Western Aphasia Battery–Revised (WAB-R) for videoconference administration.

In the study of Theodoros et al. (2008) and Hill et al. (2009), 32 patients with aphasia due to stroke or traumatic brain injury were grouped in terms of severity and were assessed both FTF and remotely on the BNT and the BDAE-3. This assessment battery targets both oral and written language, expression and comprehension (spontaneous speech, picture description, naming, repetition, auditory comprehension, reading and writing), thus the computerized version involved the online presentation of a variety of visual and oral stimuli and the recording of both verbal (words, phrases, sentences etc.) and non-verbal responses (while using a touch screen). Overall, test scores obtained from the two delivery modes were comparable within each aphasia severity level. Additionally, the majority of participants were comfortable with the online administration process, confident with the results obtained, and equally satisfied with either FTF or web-based delivery. Researchers reported that severity of aphasia might have influenced the ability to assess two of the eight subtest clusters in the online condition, namely the naming cluster and the paraphasia tally cluster; the latter is based on the quantity and type of paraphasic errors. However, these clusters also displayed a high level of agreement between the FTF and the remote method for all severity levels. Clinical comments also indicated that the remote administration of some subtests was more laborious when assessing patients with severe aphasia. The authors concluded that although aphasia severity may increase the challenges of remote assessment, it does not have an impact on assessment accuracy.

In Palsbo's (2007) randomized agreement study, 24 poststroke patients were assigned to either a remote or a FTF assessment of functional communication based on a subset of the BDAE i.e., the subsection of *Conversational and Expository Speech* (description of the Cookie Theft picture) and the sections of *Auditory Comprehension* (commands and complex ideational material, respectively). These tasks involved the use of visual and oral stimuli, and the recording of both verbal and non-verbal responses of patients. Overall, it was found that remote assessment of functional communication was equivalent to FTF administration; percentage agreement within the 95% limits of agreement ranged from 92 to 100% for each measure of functional communication. However, it should be pointed out that percentage of exact agreement between clinicians was much lower when the BDAE was administered remotely, than when it was administered by the FTF examiner. Given that the authors did not randomize the clinicians between remote and FTF assessments, it is not clear whether this discrepancy was related to the remote administration.

In the study by Newton et al. (2013), 15 patients with aphasia were assessed in three conditions, FTF or remotely, with and without the presence of a clinician, on two language comprehension tasks, i.e., a sentence-to-picture matching task and a grammaticality judgment task, that required oral and/or visual stimuli but non-verbal responses. PWA also expressed their perceptions of each condition *via* questionnaire rating scales. High correlation of the test scores across the three conditions was attested, which suggests the remote test format was sensitive to the same factors and measured the same constructs as the FTF test version. However, it was also found that computerized administration could increase test difficulty, given that participants performed significantly lower on the remote test condition. Overall, PWA preferred the FTF assessment method, although some participants felt comfortable with the remote administration. The authors conclude that remote testing can be used for the assessment of PWA, but comparison between scores obtained by remote and FTF methods should be exercised with caution.

Dekhtyar et al. (2020) compared in-person vs. remote administrations of the WAB-R, a comprehensive test that is often considered a core outcome measure for language impairment in aphasia (Wallace et al., 2019), in 20 PWA with a variety of aphasia severities. Despite the presence of some performance inconsistencies attributed to individual variability (five of the 20 participants showed changes in aphasia classification; however, this was a result of minimal changes of the actual scores), there were no significant differences between the FTF and online conditions for the WAB-R scores, and high participant satisfaction was reported for the videoconference administration. The authors concluded that the two methods of administration of the WAB-R test can be used interchangeably.

Apart from the above attempts to validate teleassessment in aphasia, Choi et al. (2015) and Guo et al. (2017) developed tablet-based aphasia assessment applications based on conventional evaluation protocols. Choi et al. (2015) developed a mobile aphasia screening test (MAST) designed as an iPad application, based on a conventional, widely used screening, K-FAST (Ha et al., 2009), the Korean version of the Frenchay Aphasia Screening Test. Sixty stroke patients, 30 with and 30 without aphasia were assessed FTF using K-FAST and the Korean version of WAB, and remotely using MAST. MAST uses a word-to-picture

matching task to assess auditory comprehension, as well as a picture description and a phonemic verbal fluency task to assess verbal expression. Patient responses are stored in a central web portal accessible to the service providers. The system scores the comprehension task automatically, whereas the verbal expression section is scored manually offline. The authors found that MAST had high diagnostic accuracy and correlated significantly with the conventional test and screening.

Going beyond aphasia screening, Guo et al. (2017) developed and validated "Access2Aphasia", a tablet videoconferencing application for the remote comprehensive assessment of aphasia at the impairment, activity and participation levels, based on the ICF framework (World Health Organization, 2001). Thirty PWA were randomized into either FTF or remote administration of the spoken word to picture matching and the naming tasks of the Psycholinguistic Assessment of Language Processing in Aphasia (PALPA), and the Assessment of Living with Aphasia (ALA) questionnaire. The study found moderate to almost perfect agreement of the online and the conventional assessment, and comparable intra- and inter-rater reliability for the two conditions.

Regarding specifically the remote elicitation and analysis of narrative discourse, Brennan et al. (2004) and Georgeadis et al. (2004) assessed 40 patients with a recent onset of either a stroke or TBI on the production and comprehension of spoken narratives, both in-person and by videoconference. The authors used a standardized discourse elicitation protocol, the Story Retell Procedure (Doyle et al., 2000). In each condition, patients listened to three pre-recorded stories accompanied by a series of black-and-white line drawings, that, in the remote condition, were scanned and displayed on a computer monitor. After the completion of the story, all pictures were displayed together, and the clinician asked each participant to retell the story using her/his own words. In both conditions, the patient's narrative was recorded and analyzed offline, in terms of the percent of information units, i.e., percent of intelligible utterances that convey accurate information relevant to the story (McNeil et al., 2001). The researchers did not report any significant differences in the patients' performance between the two assessment conditions. Additionally, a high level of acceptance of the remote version of the narrative elicitation procedure was reported. However, it is worth mentioning that patients with TBI were less likely, compared to stroke patients, to use videoconferencing for communication with the clinician. Very recently, AphasiaBank has also released an electronic version of the AphasiaBank standardized discourse elicitation protocol, which can be used for assessing PWA remotely in the context of a videoconference.[3]

The above literature review underscores the continuous and increasing need for developing and testing novel, remote assessment methodologies across all modalities and domains of communication, addressed to PWA of different types and severity levels. The available body of research has investigated web-based versions of conventional Aphasia tests and tools that specifically elicit and analyze narrative discourse. All past research studies support the validity, feasibility and reliability of web-based assessment for PWA and indicate that conventional assessment procedures can be modified to accommodate computer delivery. Although most discrepancies found between the two modes of administration were minimal and non-systematic, there is some evidence, from both neurotypical populations (Noyes and Garland, 2008) and PWA (Newton et al., 2013), to indicate a small systematic disadvantage of scores obtained *via* computer-based and/or remote assessment. Although this discrepancy seems to concern more performance on standardized tests and test tasks rather than production of narrative discourse, it does imply that caution should be exercised when comparing scores collected *via* different methods of elicitation, i.e., FTF vs. remote assessment.

2. A method for online narrative discourse data collection

The presented method was designed as a result of the COVID-19 pandemic restrictions, with the aim to enable spoken narrative discourse data collection in both in-person and remote settings from PWA and neurotypical adults, using a web-based application. The speech samples collected with the presented method are intended to be used as evaluation data for a machine learning algorithm aiming to predict aphasia severity level on the basis of several linguistic features.

2.1. Protocol for narrative discourse elicitation

Since it is generally accepted that different discourse elicitation methods may impose different cognitive and linguistic demands, the literature suggests variety in discourse elicitation methods to address the diversity of clinical characteristics in aphasia (see Section 1.3). To address the issues of task variety and sufficient sample size per participant, we implemented a protocol for eliciting narrative discourse which comprises four discourse elicitation methods: (i) free narrative production (personal narrative), (ii) story production based on a single picture or a picture sequence (picture-elicited narrative), (iii) familiar story telling, and (iv) retelling of a previously heard story. These discourse elicitation methods are represented in seven narrative tasks. Four of them are adopted from the Kakavoulia et al. (2014) narrative elicitation protocol developed for the assessment of Greek-speaking PWA (Varlokosta et al., 2016), which includes the following tasks:

A. "*Stroke story*": PWA narrate the personal story of their stroke, while unimpaired participants narrate a health-related incident about themselves ("health or accident story").
B. "*the Party*": Story production based on a six-picture sequence. The picture stimuli are original and depict an adult every-day life incident: a young man, disturbed by the noise caused by a party in the adjoining apartment, gets upset and visits his neighbors to complain about it.
C. "*the Ring*": Retelling of an unknown recorded story, supported by a five-picture series. The story is original, with the structure of a traditional fairy tale. It is about the love story of a prince and a

3 The scenarios for the remote administration of the AphasiaBank protocol to PWA and unimpaired controls are available at the AphasiaBank website (Sections 1–4), https://aphasia.talkbank.org/protocol/english/.

young woman, that is hindered by the prince's evil stepmother who steals a ring, the evidence of the prince's love for the young woman. The participants are offered visual support by a sequence of five pictures depicting main events of the story while narrating the story.

D. "*Hare and Tortoise*": Retelling of a familiar recorded story, which is an adaptation of the Aesop's fable, without visual support.

Additionally, three tasks, shown below, are adopted from the AphasiaBank standard discourse protocol, to achieve the collection of larger language samples from each participant and to be compatible with the elicitation methodology of the AphasiaBank database and comparable to the studies conducted with this methodology. The two protocols share the task of the free production of a personal story (*Stroke story*).

E. "*Refused umbrella*": Story production based on a six-picture sequence.
F. "*Cat rescue*": Story production based on a single picture.
G. "*Cinderella*": Narration of the traditional Cinderella fairytale. Participants first go through the wordless storybook to refresh their memory and afterwards they narrate the story without looking at the pictures.

The AphasiaBank protocol has been widely used for the collection of spoken discourse from PWA and healthy controls. The repository hosts data from nearly 300 PWA and 200 age-matched unimpaired controls (MacWhinney and Fromm, 2016). The AphasiaBank data are being analyzed with CLAN programs and have been used for the investigation of several research topics addressing PWA and unimpaired controls, such as discourse, grammar, gesture, lexicon, fluency, automatic classification, social factors, and treatment effects. Thus, it comprises tasks that have been multiply validated across different types of aphasia at different levels of analysis. The Greek protocol for oral narrative discourse collection has been used to collect data from 22 PWA and 22 age and education-matched controls, the spoken discourse samples of whom comprise the Greek Corpus of Aphasic Discourse (GREECAD) (Varlokosta et al., 2016). Table 1 presents the correspondence between discourse elicitation methods and narrative tasks of the implemented protocol.

The *Party*, the *Cat rescue* and the *Refused umbrella* tasks involve narration on the basis of picture stimuli. However, *Cat rescue* involves a single picture stimulus, and, therefore, differs from *the Party* and the *Refused umbrella* tasks, which involve a six-picture sequence stimulus. The *Party* picture sequence depicts an incident of adult everyday life, compared to the *Refused umbrella* which represents a child life incident. The *Ring* and *Hare and tortoise* represent two retelling tasks, since they require the reproduction of a recorded story, but differ significantly in the linguistic and cognitive demands they impose to participants. The *Ring* is an unknown lengthy story, with many episodes and a complex plot, characteristics which increase the linguistic and cognitive demands of the task. The visual support offered by five pictures, which depict the main events of the story, is expected to compensate for the increased task demands. Hare and tortoise is a familiar story, especially in the Greek culture, thus no visual support is offered. Regarding the *Cinderella* narration task, several terms have been used to describe it. It has been termed as "story narrative" (MacWhinney et al., 2011), "story retelling" (Stark and Fukuyama, 2021) and "storytelling" (Fromm et al., 2022). We chose not to refer to this task as "retelling", since it does not require reproduction of an already heard story, but as "familiar storytelling", to indicate the activity of the narration of a well-known fairytale.

2.2. Web application for data collection

The narrative discourse elicitation protocol presented in Section 2.1., including task instructions, as well as visual and audio stimuli for each task, is integrated in a custom web application for use on computer. The web application development had to meet specific requirements regarding different aspects of use: (a) consistent administration, (b) secure audio recording and good sound quality, (c) friendly and easy-to-use interface.

The parameter of consistency of administration is associated with the specifications regarding the order of task presentation, timing and modality of task stimuli, as well as content and modality of task instructions, which should follow the same principles in both the pen-and-paper and online administration, ensuring that the elicitation of spoken discourse samples is representative of the participants' language abilities, and that the elicitation is carried out in a consistent way across investigators. Presentation of tasks follows a linear sequence, beginning with the AphasiaBank tasks (A, E, F, G), followed by the Greek elicitation protocol tasks (B, C, D). Task administration follows a consistent flow, starting with task instructions, which are presented in both written and audio format to ensure that the goal of the task is clear to the participant. Subsequently, the task stimuli are presented, according to each elicitation method. In narrative tasks prompted by a picture or a picture sequence (tasks B, E, F), the picture(s) is/are presented to the participant on screen. In the case of picture sequences, the pictures appear all at once, with a number indicating their order (Figure 1). The order of pictures is pointed out by the investigator, who allows the participant enough time to go through the picture sequence and form a mental representation of the story. The pictures appear in their actual size or can be enlarged (e.g., *the Party*), so the participant can examine them in detail, serially, using the "next" navigation arrow (Figure 2).

In the case of the *Cinderella* task, the picture book has been integrated in its digital format, following the exact pagination of the printed book version. The participant can examine the book linearly, flipping through its pages back and forth, using the navigation arrows.

In tasks requiring the retelling of a recorded story (tasks C and D), story playback is initiated and controlled using an integrated audio player. In case retelling is supported by pictures (task C), the pictures appear on screen while the participants are listening to the recorded story; picture enlargement option is also available (Figure 3).

Once stimuli presentation is completed, voice recording begins using the microphone icon, which appears on each task screen at the same position. Recording ends by pressing the microphone icon again and it is submitted by pressing the "next task" icon.

TABLE 1 Correspondence between narrative tasks and narrative discourse elicitation methods.

Narrative tasks	Discourse types			
	Personal narrative	Picture-elicited narrative	Familiar storytelling	Story retelling
Stroke/health story	X			
the Party		X		
the Ring				X
Hare and tortoise				X
Cat rescue		X		
Refused umbrella		X		
Cinderella			X	

FIGURE 1
Home screen of the *Refused umbrella* task (picture adopted from the AphasiaBank protocol, MacWhinney et al., 2011).

2.3. Administration specifications

An elaborated Elicitation and Administration Guide[4] is provided to the investigators who administer the protocol. The guidelines serve the following main goals: (a) to ensure elicitation of the narrative discourse type, and not descriptive or conversational discourse, (b) to minimize verbal interventions from the part of the investigator, so the audio files acquired will be as "free" of non-participants' voice as possible, and (c) to facilitate uniformity of administration across settings and investigators and, therefore, acquisition of comparable speech samples across participants. The guidelines follow the AphasiaBank instructions[5] and include

4 The Elicitation and Administration Guide is available in Greek at https://www.planv-project.gr/files/applications/Elicitation_Administration_Guide_PLan-V.pdf.

5 Instructions for the remote and local administration of the AphasiaBank protocol are available at https://aphasia.talkbank.org/protocol/english/.

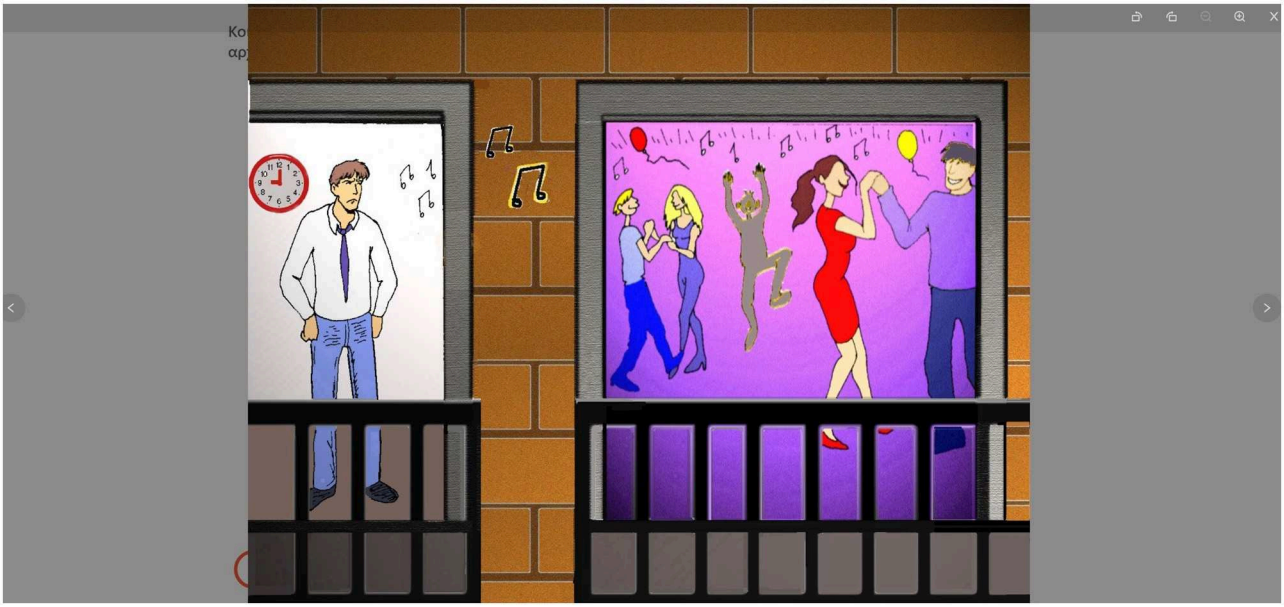

FIGURE 2
Picture enlarging functionality of the *Party* task (picture adopted from the Kakavoulia et al., 2014 protocol).

specific prompts that should be used in case there are difficulties in story production, as well as troubleshooting questions for each task.

The instructions of the seven discourse tasks emphasize the production of the narrative mode and prompt participants toward the production of stories with a beginning, middle and end (Wright and Capilouto, 2009). At the same time, specific instructions are given regarding verbal encouragements and interruptions. Investigators are instructed to avoid verbal encouragements and use non-verbal cues instead, such as facial expressions, eye contact and gestures. Moreover, type and degree of verbal encouragements and facilitating questions are controlled, allowing for gradually more specific prompts, which range from general encouragements to story-specific aids, in case of no response or serious speech halting.

The administration procedure is preceded by a brief introduction about the data collection process and its purpose. Participants are requested to narrate the stories at their own pace, and it is clarified that the investigator will not intervene, unless it is needed. They are also informed about issues of personal data protection and data access rights.

Participants sit in front of a computer screen or a laptop, which has a built-in microphone. In case there is an external microphone, it is placed close to the participants, between them and the screen. Investigators are advised not to stand next to the participants while they are examining the picture stimuli and producing the stories, so no shared knowledge of the story is assumed by the participant, a factor which might affect the linguistic characteristics of the narrative (Holler and Wilkin, 2009). For this reason, investigators are facing the participant and are not sitting next to her/him. Since the elicitation protocol is lengthy, the web application design allows its administration in several sessions. When a session of a number of tasks is completed, the investigator can exit the application, while all data collected are saved in the database. Login to the application with the same account initiates a new session, where the investigator can select the next task from a dropdown menu, skipping the tasks that have been already completed.

The same procedure, as well as elicitation and administration guidelines, are being followed in case a participant is assessed remotely in her/his own environment. Remote administration is carried out *via* the web-based application in the context of a teleconference, while the investigator is offering guidance to the participant for navigating through the application (see more in Section 3.2). Verbal encouragements and interruptions are still recommended to be avoided, as in face-to-face sessions. However, they are less likely to occur in remote elicitation settings, since the investigators are advised to keep their microphone muted during participants' recording. Moreover, the factor of spatial proximity between the participant and the investigator while narrating a story, which might favor the assumption of shared knowledge of the story, is not present in the remote elicitation setting. It should also be noted that full implementation of remote administration is feasible with the use of any remote desktop software, if desired, which allows the investigator greater control of the application operation and minimizes participants' interactions with the computer.

2.4. Technical specifications

The application is written in Javascript, using the ReactJS environment for the development of the interactive interface with browser access. Audio files and images are uploaded and stored on a remote computer (server), which uses Flask server software, while user accounts are stored in databases. Development issues were primarily related to the quality of sound recording, which is essential for collecting samples for speech processing purposes,

FIGURE 3
Screen of the *Ring* task, with audio and visual stimuli (picture adopted from the Kakavoulia et al., 2014 protocol).

especially from people whose speech impairments might have affected their voice quality in terms of intensity. More specifically, the AudioWorklet technology was used, so that the recording and any other processing can be done in a separate calculation thread, leaving the basic functions of the user's computer unaffected. This ensures there are no distortions or interruptions in the audio files. The files are recorded in high quality, with a sample rate of 44,100 Hz, 16 bit (cd quality), and sent *via* streaming during the recording to ensure a smooth user experience. Recordings are saved in a password-protected database, which provides data tracking information for user, task, session, date and time.

3. Evaluation of the web-based data collection method in neurotypical adults

The web application for narrative discourse data collection has so far been implemented for collecting narratives from PWA in an in-person setting, where SLTs have been administering the protocol of narrative tasks in the context of patients' in-house treatment. Moreover, spoken narratives are being collected from unimpaired individuals, both in in-person and remote settings. Data are being collected, manually transcribed and processed with Natural Language Processing (NLP) tools, with the aim of quantifying text properties of spoken discourse production at several types and degrees of impairment.

To evaluate and investigate the efficacy of the presented web-based data collection method in remote elicitation settings, a pilot study was conducted, aiming to compare narratives elicited remotely to narratives collected in an in-person elicitation mode from unimpaired adults. The main research question of the study is: *Are spoken narratives produced in remote elicitation settings comparable to the ones produced in in-person settings, in terms of specific characteristics of language production at the discourse level*?

3.1. Participants

The study involved ten unimpaired, Greek-speaking adult participants (five male and five female) and applies a random sampling method. However, since the objective of the study is to investigate the efficacy of an online discourse elicitation method, which addresses a specific adult population, PWA, certain

TABLE 2 Participants' demographics and cognitive tests scores.

Sex (*N*)	Age (years)	Education (ISCED level)	MMSE score	5-objects score
	Mean (SD) Min–Max		Mean (SD) Min–Max	Mean (SD) Min–Max
F (5)	65.8 (4.08)	4–5	29.2 (0.74)	25 (0.00)
	61–70		28–30	−25
M (5)	61.8 (5.06)	6–8	30 (0.00)	25 (0.00)
	57–70		−30	−25

inclusion criteria have been applied, so the sample is comparable to the characteristics of the population that the method intends to assess. Therefore, since aphasia is more common in older than younger adults[6] (Ellis and Urban, 2016), the participants' age ranges from 57 to 70 years (mean age = 63.8). Moreover, since it is evidenced that educational level affects language abilities of PWA (González-Fernández et al., 2011), as well as of unimpaired individuals (Radanovic et al., 2004), the independent variable of educational level should be controlled. Therefore, the sample involves participants with a minimum ISCED level 4, i.e., individuals who have completed upper secondary education ($N = 3$), participants having completed short-cycle tertiary education (ISCED level 5, $N = 2$), as well as participants who hold a Bachelor's, Master's, or Doctoral degree (ISCED levels 6, 7, 8, $N = 2$, 1, and 2, respectively), according to the International Standard Classification of Education (UNESCO-UIS, 2012). All participants were screened using neuropsychological tools, namely the Mini Mental State Examination (MMSE) (Folstein et al., 1975) and the 5-Object cognitive screening test (Papageorgiou et al., 2014), to ensure that their cognitive abilities were within the norms. Participants' demographic details are presented in Table 2.

The study was approved by the Bioethics Committee of the coordinating organization, University of Patras, Greece. All participants have signed a Participation Consent Form, after being informed on the study's aims and objectives *via* a Participant Information Sheet.

3.2. Data collection procedure

The study uses a within-subjects experimental design, since all participants took part in both experimental conditions: in-person and remote elicitation set up. The same investigator administered the tasks to the same participant in both conditions, to eliminate the effect of the investigators' communication style on participants' language production. Investigators were two linguists and one SLT, members of the research team, experienced in administering assessment protocols and familiar with the specific narrative elicitation tasks. To eliminate the effect of prior knowledge of the stories and task familiarity on language production, as a result of the administration order between the two conditions, the sample was split into two groups, with the first group being first investigated in the remote condition and the second in the in-person condition. The two sessions had a minimum time distance of 1 week from each other.

In the remote elicitation setting, participants performed the narrative elicitation tasks *via* the web application in their own environment, with the assistance of an investigator in the context of a virtual meeting (video conferencing). The participants shared their screen, so the investigator could help them enter their account credentials and navigate through the application. The investigators followed the same guidelines regarding order of tasks, instructions, and interventions. Recording was directly done through the participants' laptop built-in microphone, and not through the investigators' laptop speakers, to avoid sound distortion. Investigators had their microphone muted, to avoid voice interference and overlaps during the participants' narration.

In the in-person elicitation setting, the participant was in the same room with the investigator, who administered the tasks *via* the web application following the same guidelines.

3.3. Data transcription and processing

A total of 139 spoken narrative samples was collected, 70 elicited in the in-person condition and 69 in the remote condition. Ten participants performed seven narrative tasks each, in two conditions, with one missing data-point for one narrative task in the remote condition. The recorded narratives were manually transcribed in an orthographic format by three researchers using the ELAN software. Transcriptions were manually time-aligned, to allow for the automatic calculation of duration, and manually segmented into utterances, following the AphasiaBank guidelines for utterance segmentation; each utterance includes only one main clause along with its depended subordinate clauses. Repetitions, reformulations, and false starts were not included in utterances transcription for uniform word count and MLU calculation. According to the AphasiaBank guidelines, the period and the question mark were used as utterance terminators. Also, wide use of comma was applied, to indicate boundaries of clauses and phrases, that would facilitate NLP tools to perform accurate syntactic parsing. All transcripts were evaluated and normalized by a single researcher to ensure uniformity in utterance segmentation, as well as consistent application of orthographic and punctuation criteria (e.g., use of comma before a subordinate clause, use of full stop only at utterance final position, word contractions etc.). Table 3 presents an overview of the study dataset.

6 National Institute of Deafness and Other Communication Disorders. Available online at: https://www.nidcd.nih.gov/health/aphasia (last accessed April 6, 2022).

TABLE 3 Overview of the study dataset.

	In-person	Remote
Narrative samples (*N*)	70	69
Total *N* of tokens	14,137	14,101
Mean *N* of tokens per participant (SD)	1,411.9 (424.57)	1,452.58 (423.006)
Min–Max *N* of tokens per participant	918–2,441	936–2,492
Total duration (s)	7,827.86	7,817.004
Mean duration per participant(s) (SD)	782.24 (206.38)	783.63 (214.86)
Min–Max duration per participant(s)	595.99–1,218.34	561.54–1,292.38

Transcripts were extracted by participant and task in plain text format and processed with the Neural NLP Toolkit for Greek[7] (Prokopidis and Piperidis, 2020). The Neural NLP Toolkit for Greek is a state-of-the-art suite of NLP tools for the automated processing of Greek texts, developed at the Institute for Language and Speech Processing/Athena Research Center (ILSP/ATHENA RC). It currently integrates modules for part of speech (POS tagging), lemmatization, dependency parsing and text classification. The toolkit is based on code, models and language resources developed at the NLP group of ILSP.

3.4. Measures of narrative discourse production

According to recent literature reviews (Bryant et al., 2016; Pritchard et al., 2018) on the analysis of discourse in aphasia, more than 500 linguistic variables are being used to measure spoken language abilities and intervention outcomes of PWA. To address the variety and heterogeneity in methods, measures and analyses of spoken discourse samples, recent research initiatives are being undertaken toward the standardization of measures and methods (Stark et al., 2021b), as well as the identification and evaluation of primary linguistic variables for the reliable assessment of language abilities in aphasia across discourse types and elicitation methods (Stark, 2019). Moreover, given the growing availability of shared databases, such as AphasiaBank, as well as of tools for automated language analysis, statistical and machine learning methods are being increasingly applied for the automatic analysis, assessment and classification of PWA's speech samples, quantifying their linguistic properties and translating them into features used for the computational modeling of aphasia (Stark and Fukuyama, 2021; Fromm et al., 2022).

Following Stark (2019), who extracted from AphasiaBank data a set of eight primary linguistic variables which serve as proxies for various language abilities at spoken discourse level, the same set of features was selected to measure spoken language production of participants at both experimental conditions. These features correspond to the language abilities of linguistic productivity (MLU), content richness (propositional density), fluency (words per minute), syntactic complexity (verbs per utterance, open/closed class words, noun/verb ratio), lexical diversity (lemma/token ratio) and gross linguistic output (number of words). These linguistic variables have been evaluated by Stark (2019) in a large sample of PWA and unimpaired controls, drawn from the AphasiaBank database, across three discourse types, expositional, narrative, and procedural discourse, corresponding to four discourse elicitation tasks of the AphasiaBank protocol: *Broken window, Cat rescue*, *Cinderella* and *Peanut Butter and Jelly* (procedural discourse). Her analysis showed significant effects of discourse type on the linguistic properties of spoken discourse in both groups, with similar findings across groups regarding discourse type sensitivity to primary linguistic variables.

In the present study, the measure representing lexical diversity was modified, since lemma/token ratio measure, which considers inflectional variants of the same lemma as the same type, was favored over the most commonly used type/token ratio measure, which treats inflected forms of the same lemma as different types. This decision was based on studies of lexical diversity in narrative discourse of PWA and unimpaired controls (Fergadiotis and Wright, 2011; Fergadiotis et al., 2013) which performed a lemma-based analysis of lexical diversity. In these studies, different inflected forms of the same word, for example eat, eats, ate, were counted as one and the same type. The reason for counting only unique lexical representations as separate types was to avoid conflating the measure of lexical diversity with the one of grammaticality, as reflected on the use of different inflected forms of the same lemma. Lemma/token ratio has also been applied in measuring lexical diversity in EFL learner corpora (Granger and Wynne, 1999), as the use of different lemmas (such as go, come, leave, enter, return) indicates greater lexical richness than the use of different inflected forms of the same lemma (such as go, goes, going, gone, went). For these reasons, as well as given that Greek is a highly inflected language, the present study adopts the lemma-based analysis of lexical diversity as a more representative measure of speakers' vocabulary range.

Moreover, two additional measures of syntactic complexity at the inter-clausal level were implemented, subordinate/all clauses ratio and mean dependency tree height. These measures were selected as relevant to NLP-based linguistic features extraction for the automatic processing of language data. They have been widely employed as highly effective measures of linguistic complexity in various fields of the automatic processing of texts, such as automatic text readability assessment (Vajjala and Meurers, 2012), Second Language Acquisition (SLA) research (Chen et al., 2021), automatic analysis of language production in aphasia (Gleichgerrcht et al., 2021), automatic Primary Progressive Aphasia (PPA) subtyping (Fraser et al., 2014) and automatic Alzheimer's Disease identification (Fraser et al., 2016). Subordinate/all clauses ratio represents the ratio of all subordinate clauses (complement, adverbial, relative clauses) to all clauses produced in the same narrative, including subordinate and main clauses. Mean dependency tree height measures the height of the dependency tree

7 The Neural NLP Toolkit for Greek is available as a web application at http://nlp.ilsp.gr/nws/.

10.3389/fcomm.2023.919617

TABLE 4 Linguistic measures for analyzing discourse production and their correspondence to language abilities (*a.-h.* adapted from Stark, 2019).

Measure	Definition	Language ability
a. Mean length of utterance (MLU)	Average number of words per utterance (excluding repetitions and reformulations)	Linguistic productivity
b. Propositional density	Number of verbs, adjectives, adverbs, prepositions, and conjunctions divided by the total number of words	Content richness
c. Words per minute	Total number of tokens divided by total story duration	Fluency
d. Verbs per utterance	Average number of verbs per utterance	Syntactic complexity
e. Lemma-token ratio	Number of lemmas divided by the total number of words (tokens)	Lexical diversity
f. Open-closed class words ratio	Ratio of open class words (nouns, verbs, adjectives, adverbs) divided by closed class words (all other classes)	Syntactic complexity
g. Noun-verb ratio	Ratio of nouns to verbs	Syntactic complexity
h. Number of words (tokens)	Total number of words produced	Gross output
i. Subordinate/all clauses ratio	Number of subordinate clauses (complement, adverbial, relative clauses) divided by the number of all clauses produced	Syntactic complexity
j. Mean dependency tree height	Height of the dependency tree (syntax tree)	Syntactic complexity

(corresponding to the syntax tree). The higher the dependency tree, the more complex the syntactic structure.

Table 4 presents the linguistic measures applied in the present study for the analysis of narrative discourse production and their correspondence to language abilities.

All these measures were automatically calculated using the tokenization, lemmatization, POS tagging and dependency parsing modules of the Neural NLP tool for Greek. More specifically, the measures of mean dependency tree height and subordinate clauses ratio were calculated using the dependency parser module. Since the dependency parser analyzes sentences, which consist of multiple main clauses -together with their subordinate clauses- connected with coordinating conjunctions, manual post-processing of transcripts was carried out, in order to convert utterances into sentences. As part of this process, several utterances were merged into a single sentence, using mainly intonational criteria, as shown in the following example:

> *and then the witch found a pumpkin from the garden.*
> *she gave it a hit.*
> *and she transformed it into a carriage.*
> *and the two mice that accompanied Cinderella, she transformed them into two very nice horses.*

Each one of the above lines corresponds to a separate utterance of the transcript. At post-processing, following the speaker's intonation contour, the four utterances were merged into a single sentence, beginning with a capital letter, and ending in a full stop or a question mark:

> *And then the witch found a pumpkin from the garden, she gave it a hit, and she transformed it into a carriage, and the two mice that accompanied Cinderella, she transformed them into two very nice horses.*

Post-processing was carried out by a single researcher, to ensure uniformity of sentence segmentation. Extensive evaluation of the automatically computed feature values was performed by two researchers, which led to script modifications, until minimal calculation errors were identified. For example, there were cases of passive participles tagged as verbs by the POS tagger, on the basis of their morphological characteristics, but actually had the syntactic role of an adjectival modifier (example 1) or a nominal subject (example 2) or object (example 3). The script written to calculate POS from the POS tagger output was modified to assign the POS tag "adjective" (example 1) and "noun" (examples 2–3) to the respective words.

(1) *Η καημένη* (POS: VERB | VerbForm: Part | syntactic role: amod) *η Σταχτοπούτα τα έβλεπε όλα αυτά.* (*Poor Cinderella was watching all this.*)
(2) *Στη γιορτή αυτή υπάρχουν διάφοροι καλεσμένοι.* (POS: VERB | VerbForm=Part | syntactic role: nsubj) (*In this party there are several guests.*)
(3) *Κι έτσι λοιπόν ο πρίγκιπας βρίσκει την αγαπημένη* (POS: VERB | VerbForm=Part | syntactic role: obj) *του.* (*And so, the prince finds his beloved.*)

4. Results

As a first step of analysis, correlations between the linguistic measures presented in Section 3.4 were calculated, to explore whether the selected measures contribute significantly in describing the language abilities of the participants at the discourse level. Linguistic variables were averaged across all narrative tasks and across the two elicitation conditions, in-person and remote. Results are presented in Table 5.

The results indicate some strong and expected correlations between certain linguistic variables. MLU correlates with three variables of syntactic complexity, i.e., verbs per utterance, ratio of dependent to all clauses, and mean tree height, indicating the interdependence of the utterance length with the structural complexity at utterance and sentence level. Accordingly, a strong correlation is found between the aforementioned variables of syntactic complexity, indicating that there is a close two-way relationship between the number of verbs in an utterance, the density of subordination at inter-clausal level, and the overall structural complexity at sentence level, as represented by the syntax tree. These results reveal some predictable but meaningful interactions between variables of linguistic productivity (MLU) and syntactic complexity (VPU, MTH, and Dep/Cl).

TABLE 5 Correlations between linguistic variables of narrative discourse production across elicitation condition.

	MLU	WPM	VPU	TTR	NVR	NoW	OCCW	Dep/Cl	MTH	PrD
MLU	1									
WPM	−0.470	1								
VPU	**0.981**	−0.412	1							
LTR	0.487	−0.477	0.410	1						
NVR	0.656	−0.642	0.538	0.762	1					
NoW	−0.195	0.382	−0.110	−0.710	−0.461	1				
OCCW	−0.343	0.078	−0.267	−0.165	−0.269	0.312	1			
Dep/Cl	**0.952**	−0.486	**0.958**	0.541	0.676	−0.167	−0.192	1		
MTH	**0.987**	−0.496	**0.968**	0.496	0.692	−0.169	−0.335	**0.952**	1	
PrD	−0.581	0.545	−0.519	−0.829	−0.890	0.551	0.243	−0.662	−0.638	1

MLU, mean length per utterance; WPM, words per minute; VPU, verbs per utterance; LTR, lemma-token ratio; NVR, noun-verb ratio; NoW, number of words; OCCWR, open-closed class words ratio; Dep/Cl, dependent clauses divided by total number of clauses; MTH, mean tree height; PrD, propositional density. Values in bold indicate strong correlations ($> \pm 0.95$).

We subsequently analyzed the linguistic variables under investigation across the two experimental settings, and across the seven narrative tasks. As described in Section 3.2, the two experimental groups consisted of the same participants. Therefore, these two samples are considered as dependent, which entails paired measurements of the same participant. Subjects within each group are independent. Given the small sample size (<30), it is hard to test the sample data for normality. Therefore, the non-parametric equivalent of the paired *t*-test, the Wilcoxon signed-rank test, was conducted. Supplementary Table 1 presents the results of the dependent-sample Wilcoxon test conducted for each one of the ten variables under investigation per narrative, making a total of 70 tests. Formally, the test hypothesis is formulated as follows:

(1) Null hypothesis (H0): The difference between the pairs follows a symmetric distribution around zero.
(2) Alternative hypothesis (HA): The difference between the pairs does not follow a symmetric distribution around zero.

The results did not reveal any statistically significant differences (p-value > 0.05) for 66 out of the 70 tests. The fact that no statistically significant differences were demonstrated for the vast majority of the linguistic variables analyzed per narrative task suggests that the administration condition did not have a significant effect on the linguistic properties of the spoken narratives produced by the study participants. In four cases, statistically significant mean differences emerged, i.e., in MLU, Verbs per Utterance and Mean Tree Height for *Cinderella,* and in Noun-Verb ratio for *the Ring* task. In the case of the *Cinderella* task, the means of MLU, VPU and MTH were significantly higher in the remote elicitation condition, as was the mean of NVR in the case of the *Ring*. All these measures correspond to features of syntactic complexity, while three of them, MLU, Verbs per Utterance and Mean Tree Height, were strongly correlated variables (Table 5). The *Cinderella* task is found to be one of the most sensitive discourse elicitation tasks to measure propositional density and syntactic complexity in spoken narratives of both PWA and unimpaired controls (Stark, 2019). Although this may not directly serve as an explanation of the rejection of the null hypothesis for three syntactic complexity variables, further investigation in a larger sample of participants is needed to explore whether the remote condition elicits more syntactically complex language, at least in the case of *Cinderella*, which has been demonstrated as a good elicitation task to measure syntactic complexity.

5. Discussion

This paper presents the design, methodological considerations and requirements of a web-based method designed to facilitate spoken narrative discourse data collection from Greek-speaking PWA and unimpaired adults in remote as well as in in-person administration conditions for online assessment purposes. The application comprises a 7-task protocol for narrative discourse elicitation, guidelines to ensure reliable elicitation of narrative discourse type, appropriate sampling of participants' spoken discourse, and consistent administration across investigators and settings. Transcription procedures of speech samples, as well as procedures for the automated analysis of transcripts with available NLP tools for Greek texts, enabling the calculation of linguistic features that can serve as indicators of language abilities are also described.

As a first step for validating the presented method, a pilot study was conducted including only unimpaired adults and comparative results of narrative discourse produced in two different conditions of data collection, i.e., in-person and remote setting, are presented. The aim of the study was to investigate the feasibility of the presented method to elicit spoken narratives in remote collection settings that are comparable to the ones produced in in-person settings, in terms of specific characteristics of language production at the discourse level. Spoken narratives were collected with the use of the presented application in both conditions, using a within-subjects experimental design.

A set of ten linguistic variables representing various language abilities at the spoken discourse level, i.e., linguistic productivity, content richness, fluency, syntactic complexity at utterance and sentence level, lexical diversity, and verbal output, was selected to quantify spoken language production of participants in both

experimental conditions. Statistical analysis for each variable per narrative task indicated non-significant differences for most of the paired samples mean difference measurements, a finding which indicates the efficacy of the presented method to collect spoken narrative discourse samples from neurotypical adults with similar linguistic properties in both conditions.

A major limitation of the presented study is related to the small number of participants. Even though the statistical hypotheses were tested on a sufficient dataset of speech samples per participant, in terms of task variety, time duration and number of tokens (Brookshire and Nicholas, 1994; Boles and Bombard, 1998), the small sample size compromises the generalizability of results to a broader population of neurotypical adults.

Moreover, the fact that the method's validation included only unimpaired controls, and not PWA, does not allow considering the present study results generalizable to the population of PWA. PWA represent a vulnerable population that often suffers from coexisting chronic physical dysfunctions, cognitive impairments as well as negative social and emotional outcomes, such as depression and low social participation (Kauhanen et al., 2000; Mayo et al., 2002; Hilari et al., 2015). These conditions might affect PWA's ability to use technology-enhanced assessment environments remotely, without in-person supervision by an investigator. Despite the fact than none of the study participants needed help in navigating through the application or in setting up the teleconference, replicating the study with PWA, or even with unimpaired participants of lower educational level, might reveal issues related to the independent use of technology that did not occur in the present study.

Therefore, the planned next step for validating the presented method is to conduct the same study on participants with aphasia, to demonstrate its feasibility in collecting comparable data from adults with language impairments in remote and in-person assessment settings. This study can include participants of different aphasia severity levels, with the aim to identify possible differences in linguistic variables of spoken discourse across elicitation conditions and to investigate the effect of aphasia severity on these differences. Moreover, evaluation of the selected linguistic variables can identify effects of narrative tasks on language production properties in each population, PWA and neurotypical adults, and in comparison with each other, which can be further explored across elicitation condition, i.e., remote and in-person. A complementary area of future research is to investigate the contribution of additional linguistic variables, such as features related to informational content and narrative macrostructure, in the description of language abilities of both PWA and unimpaired adult populations, as well as the effect of elicitation condition on these variables.

Future work will also involve replication of the study in a larger sample of unimpaired adults of different age groups and educational levels, with the aim to investigate the impact of these demographic variables on different elicitation conditions. Age and educational level might have an effect on participants' performance in the remote elicitation condition, which is heavily related on technology skills, so future research could contribute to testing this hypothesis.

Given the above limitations, our findings are aligned with prior studies (Palsbo, 2007; Theodoros et al., 2008; Hill et al., 2009; Dekhtyar et al., 2020) which compare online assessment methods addressed to individuals with language and communication disorders in remote and in-person conditions, suggesting that both settings produce comparable results in terms of language production. Concerns raised regarding participants' technology skills need to be considered for the method's effective implementation, by adding special instructions for participants and troubleshooting guidelines for investigators in case of remote elicitation of spoken discourse from either impaired or unimpaired individuals. Instructions can include issues such as setting up a teleconference, screen sharing, microphone muting and unmuting, as well as navigating through the application. However, it is worth noting that the use of the presented web-based application is still feasible even in case of participants with limited or no technology skills, with the use of a remote desktop software by the investigator, which will allow full control of the participant's computer.

The presented web-based data collection method is currently being employed for collecting spoken language data from PWA in-person and from unimpaired individuals, either remotely or in-person. This dataset (speech samples, transcripts, features measured and labels for aphasia class) serves as a golden corpus, which provides the ground truth, on the basis of which a machine-learning system for the automatic classification of aphasia in Greek will be assessed and evaluated. A substantial amount of manual work is carried out for compiling this corpus; manual transcription and time-alignment, utterance segmentation and sentence splitting. Since a large amount of data is required to train accurate linguistic models for automatic classification purposes, ongoing research activities are being carried out that aim to automate manual work involved in the data processing and analysis pipeline (Chatzoudis et al., 2022) for aphasia classification purposes, such as Automatic Speech Recognition in Greek for transcription and time alignment.

In sum, the presented method, as evidenced from the present study findings, offers an applicable, feasible and valid framework for both in-person and remote online elicitation of spoken narrative discourse samples for the assessment of language abilities of adult populations without language disorders. The next step will be to investigate its feasibility to collect comparable spoken discourse data from adults with language disorders in remote and FTF settings. In line with the current literature on language and communication disorders assessment and intervention, which highlights the need for the modification of conventional pen-and-paper methods to accommodate technology-enhanced tools and applications, the present paper provides some initial evidence toward the reliable implementation of technology applications for remote language data collection and assessment of language skills.

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 study involving human participants was reviewed and approved by the Bioethics Committee, University of Patras, Patras, Greece. The participants provided their written informed consent to participate in this study.

Author contributions

SS, AK, and VK: conceptualization. SS, MN, IP, A-LD, AK, SV, and AT: methodology. GC, MP, AK, and VK: software. SS, MN, and A-LD: data collection, transcription, and annotation. SS, GC, and MP: data processing and analysis. SS, MN, and IP: writing. SS, MN, IP, A-LD, SV, and AT: review and editing. AT: scientific coordination. IP: project administration. All authors have read and agreed to the published version of the manuscript.

Funding

This research has been co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T2EDK-02159, "A Speech and Language Therapy Platform with Virtual Agent" - PLan-V, PI: AT).

Acknowledgments

The authors are grateful to Dimitris Mastrogiannopoulos for his contribution to the development of the web application, Anthi Zafeiri, PhDc, for data collection, Theophano Christou, PhD, for data transcription, as well as to Dr. Vassilis Papavasileiou for his contribution to the statistical analysis of data.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2023.919617/full#supplementary-material

SUPPLEMENTARY TABLE 1
Measures of linguistic productivity, content richness, fluency, syntactic complexity, lexical diversity and verbal output, statistically compared between in-person and remotely elicited speech samples using the non-parametric Wilcoxon signed-rank test for comparison of means. *P*-values are listed, with an asterisk (*) indicating statistical significance.

References

American Speech-Language-Hearing Association (2016). *2016 SIG 18 Telepractice Survey Results*. Available online at: https://www.asha.org/siteassets/practice-portal/telepractice/2016-telepractice-survey.pdf (accessed April 13, 2022).

Armstrong, E. (2000). Aphasic discourse analysis: the story so far. *Aphasiology* 14, 875–892. doi: 10.1080/02687030050127685

Armstrong, E., Ferguson, A., and Simmons-Mackie, N. (2011). "Discourse and functional approaches to aphasia," in *Aphasia and Related Neurogenic Communication Disorders*, eds I. Papathanasiou, P. Coppens, and C. Potagas (Sudbury, MA: Jones and Bartlett Pub), 217–231.

Armstrong, E., and Ulatowska, H. K. (2007). "Stroke stories: conveying emotive experiences in aphasia," in *Clinical aphasiology: Future Directions*, eds. M. J. Ball, and J. S. Damico (New York, NY: Psychology Press).

Beeke, S., Maxim, J., Best, W., and Cooper, F. (2011). Redesigning therapy for agrammatism: initial findings from the ongoing evaluation of a conversation-based intervention study. *J. Neurolinguist.* 24, 222–236. doi: 10.1016/j.jneuroling.2010.03.002

Boles, L., and Bombard, T. (1998). Conversational discourse analysis: appropriate and useful sample sizes. *Aphasiology* 12, 547–560. doi: 10.1080/02687039808249557

Bottenberg, D., Lemme, M., and Hedberg, N. (1985). "Analysis of oral narratives of normal and aphasic adults," in *Clinical Aphasiology*, ed R. H. Brookshire (Minneapolis, MN: BRK), 241–247.

Brennan, D. M., Georgeadis, A. C., Baron, C. R., and Barker, L. M. (2004). The effect of videoconference based telerehabilitation on story retelling performance by brain-injured subjects and its implications for remote speech-language therapy. *Telemed. J. e- Health* 10, 147–154. doi: 10.1089/tmj.2004.10.147

Brookshire, R., and Nicholas, L. (1994). Speech sample-size and test-retest stability of connected speech measures for adults with aphasia. *J. Speech Lang. Hear. Res.* 37, 399–407. doi: 10.1044/jshr.3702.399

Bryant, L., Ferguson, A., and Spencer, E. (2016). Linguistic analysis of discourse in aphasia: a review of the literature. *Clin. Linguist. Phonet.* 30, 489–518. doi: 10.3109/02699206.2016.1145740

Bryant, L., Spencer, E., and Ferguson, A. (2017). Clinical use of linguistic discourse analysis for the assessment of language in aphasia. *Aphasiology* 31, 1105–1126. doi: 10.1080/02687038.2016.1239013

Chatzoudis, G., Plitsis, M., Stamouli, S., Dimou, A.-L., Katsamanis, A., and Katsouros, V. (2022). "Zero-shot cross-lingual aphasia detection using automatic speech recognition," in *INTERSPEECH 2022* (Incheon).

Chen, X., Alexopoulou, T., and Tsimpli, I. (2021). Automatic extraction of subordinate clauses and its application in second language acquisition research. *Behav. Res. Methods* 53, 803–817. doi: 10.3758/s13428-020-01456-7

Choi, Y. H., Park, H. K., Ahn, K. H., Son, Y. J., and Paik, N. J. (2015). A Telescreening tool to detect aphasia in patients with stroke. *Telemed. J. e- Health* 21, 729–734. doi: 10.1089/tmj.2014.0207

Cruice, M., Botting, N., Marshall, J., Boyle, M., Hersh, D., Pritchard, M., et al. (2020). UK speech and language therapists' views and reported practices of discourse analysis in aphasia rehabilitation. *Int. J. Lang. Commun. Disord.* 55, 417–442. doi: 10.1111/1460-6984.12528

Dekhtyar, M., Braun, E. J., Billot, A., Foo, L., and Kiran, S. (2020). Videoconference administration of the western aphasia battery-revised: feasibility and validity. *Am. J. Speech Lang. Pathol.* 29, 673–687. doi: 10.1044/2019_AJSLP-19-00023

Dietz, A., and Boyle, M. (2018). Discourse measurement in aphasia research: have we reached the tipping point? *Aphasiology* 32, 459–464. doi: 10.1080/02687038.2017.1398803

Doedens, W. J., and Meteyard, L. (2020). Measures of functional, real-world communication for aphasia: a critical review. *Aphasiology* 34, 492–514. doi: 10.1080/02687038.2019.1702848

Doedens, W. J., and Meteyard, L. (2022). What is functional communication? A theoretical framework for real-world communication applied to aphasia rehabilitation. *Neuropsychol. Rev.* 32, 937–973. doi: 10.1007/s11065-021-09531-2

Doyle, P. J., Goda, A. J., and Spencer, K. A. (1995). The communicative informativeness and efficiency of connected discourse by adults with aphasia under structured and conversational sampling conditions. *Am. J. Speech Lang. Pathol.* 4, 130–134. doi: 10.1044/1058-0360.0404.130

Doyle, P. J., McNeil, M. R., Park, G. H., Goda, A. J., Spencer, K., Lustig, A., et al. (2000). Linguistic validation of four parallel forms of a story retelling procedure. *Aphasiology* 14, 537–549. doi: 10.1080/026870300401306

Doyle, P. J., McNeil, M. R., Spencer, K. A., Goda, A. J., Cottrell, K., and Lustig, A. P. (1998). The effects of concurrent picture presentations on retelling of orally presented stories by adults with aphasia. *Aphasiology* 12, 561–574. doi: 10.1080/02687039808249558

Ellis, C., and Urban, S. (2016). Age and aphasia: a review of presence, type, recovery and clinical outcomes. *Top. Stroke Rehabil.* 23, 430–439. doi: 10.1080/10749357.2016.1150412

Fergadiotis, G., and Wright, H. H. (2011). Lexical diversity for adults with and without aphasia across discourse elicitation tasks. *Aphasiology* 25, 1414–1430. doi: 10.1080/02687038.2011.603898

Fergadiotis, G., Wright, H. H., and West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *Am. J. Speech Lang. Pathol.* 22, S397–S408. doi: 10.1044/1058-0360(2013/12-0083)

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12,189–198. doi: 10.1016/0022-3956(75)90026-6

Fraser, K. C., Meltzer, J. A., Graham, N. L., Leonard, C., Hirst, G., Black, S. E., et al. (2014). Automated classification of primary progressive aphasia subtypes from narrative speech transcripts. *Cortex* 55, 43–60. doi: 10.1016/j.cortex.2012.12.006

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis.* 49, 407–422. doi: 10.3233/JAD-150520

Fromm, D., Greenhouse, J., Pudil, M., Shi, Y., and MacWhinney, B. (2022). Enhancing the classification of aphasia: a statistical analysis using connected speech. *Aphasiology.* 36, 1492–1519. doi: 10.1080/02687038.2021.1975636

Fromm, D., Holland, A., Armstrong, E., Forbes, M., MacWhinney, B., Risko, A., et al. (2011). "Better but no cigar": persons with aphasia speak about their speech. *Aphasiology* 25, 1431–1447. doi: 10.1080/02687038.2011.608839

Georgeadis, A., Brennan, D., Barker, L., and Baron, C. (2004). Telerehabilitation and its effect on story retelling by adults with neurogenic communication disorders. *Aphasiology* 18, 639–652. doi: 10.1080/02687030444000075

Gleichgerrcht, E., Roth, R., Fridriksson, J., den Ouden, D., Delgaizo, J., Stark, B., et al. (2021). Neural bases of elements of syntax during speech production in patients with aphasia. *Brain Lang.* 222, 105025. doi: 10.1016/j.bandl.2021.105025

Glosser, G., Wiener, M., and Kaplan, E. (1988). Variations in aphasic language behaviors. *J. Speech Hear. Disord.* 53, 115–124. doi: 10.1044/jshd.5302.115

González-Fernández, M., Davis, C., Molitoris, J. J., Newhart, M., Leigh, R., and Hillis, A. E. (2011). Formal education, socioeconomic status, and the severity of aphasia after stroke. *Arch. Phys. Med. Rehabil.* 92, 1809–1813. doi: 10.1016/j.apmr.2011.05.026

Goodglass, H., and Kaplan, E. (1972). *Boston Diagnostic Aphasia Examination (BDAE).* Philadelphia, PA: Lea and Febiger.

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *Boston Diagnostic Aphasia Examination, 3rd Edn.* Philadelphia, PA: Lippincott Williams and Wilkins.

Granger, S., and Wynne, M. (1999). "Optimising measures of lexical variation in EFL learner corpora," in *Corpora Galore*, ed J. Kirk (Amsterdam; Atlanta: Rodopi), 249–257

Grimes, N. (2005). *Walt Disney's Cinderella.* New York, NY: Random House.

Guo, Y. E., Togher, L., Power, E., Hutomo, E., Yang, Y. F., Tay, A., et al. (2017). Assessment of aphasia across the international classification of functioning, disability and health using an iPad-based application. *Telemed. J. e- Health* 23, 313–326. doi: 10.1089/tmj.2016.0072

Ha, J. W., Pyun, S.-B., Lee, H. Y., Hwang, Y.-M., and Nam, K. (2009). Reliability and validity analyses of the Korean version of Frenchay Aphasia Screening Test in brain-damaged patients. *Kor. J. Commun. Disord.* 14, 46–57. Available online at: http://www.e-csd.org/journal/view.php?number=382

Harley, T. (2001). *The Psychology of Language: From Data to Theory*, *2nd Edn*. New York, NY: Psychology Press.

Hilari, K., Cruice, M., Sorin-Peters, R., and Worrall, L. (2015). Quality of life in aphasia: state of the art. *Folia Phoniatr. Logopaed.* 67, 114–118. doi: 10.1159/000440997

Hill, A. J., Theodoros, D. G., Russell, T. G., Ward, E. C., and Wootton, R. (2009). The effects of aphasia severity on the ability to assess language disorders via telerehabilitation. *Aphasiology* 23, 627–642. doi: 10.1080/02687030801909659

Holland, A. (1979). "Some practical considerations in aphasia rehabilitation," in *Rationale for Adult Aphasia Therapy*, eds. M. Sullivan, and M. Kommers (Omaha, NE: University of Nebraska Medical Center), 167–180.

Holland, A. (1982). Observing functional communication of aphasic adults. *J. Speech Hear. Disord.* 47, 50–56. doi: 10.1044/jshd.4701.50

Holler, J., and Wilkin, K. (2009). Communicating common ground: How mutually shared knowledge influences speech and gesture in a narrative task. *Lang. Cogn. Process.* 24, 267–289. doi: 10.1080/01690960802095545

Kakavoulia, M., Stamouli, S., Foka-Kavalieraki, P., Economou, A., Protopapas, A., and Varlokosta, S. (2014). A battery for eliciting narrative discourse by Greek speakers with aphasia: Principles, methodological issues, and preliminary results. *Glossologia* 22, 41–60 (in Greek). Available online at: http://glossologia.phil.uoa.gr/?q=node/107

Kauhanen, M. L., Korpelainen, J. T., Hiltunen, P., Määtt,ä, R., Mononen, H., Brusin, E., et al. (2000). Aphasia, depression, and non-verbal cognitive impairment in ischaemic stroke. *Cerebrovasc. Dis.* 10, 455–461. doi: 10.1159/000016107

Kertesz, A. (2006). *Western Aphasia Battery, Revised Edn*. San Antonio, TX: Pearson Assessment.

Labov, W. (1972). *Language in the Inner City: Studies in the Black English Vernacular*. Philadelphi, PA: University of Pennsylvania Press.

Lemme, M. L., Hedberg, N. L., and Bottenberg, D. F. (1984). "Cohesion in narratives of aphasic adults," in *Clinical Aphasiology*, ed. R. H. Brookshire (Minneapolis, MN: BRK), 215–222.

MacWhinney, B., and Fromm, D. (2016). AphasiaBank as BigData. *Semin. Speech Lang.* 37, 10–22. doi: 10.1055/s-0036-1571357

MacWhinney, B., Fromm, D., Forbes, M., and Holland, A. (2011). AphasiaBank: methods for studying discourse. *Aphasiology* 25, 1286–1307. doi: 10.1080/02687038.2011.589893

Mayo, N. E., Wood-Dauphinee, S., Côté, R., Durcan, L., and Carlton, J. (2002). Activity, participation, and quality of life 6 months poststroke. *Arch. Phys. Med. Rehabil.* 83, 1035–1042. doi: 10.1053/apmr.2002.33984

McNeil, M., Doyle, P., Fossett, T., Park, G., and Goda, A. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology* 15, 991–1006. doi: 10.1080/026870401430 00348

McNeil, M. R., Doyle, P. J., Park, G. H., Fossett, T. R. D., and Brodsky, M. B. (2002). Increasing the sensitivity of the Story Retell Procedure for the discrimination of normal elderly subjects from persons with aphasia. *Aphasiology*.16, 815–822. doi: 10.1080/02687030244000284

McNeil, M. R., Sung, J. E., Yang, D., Pratt, S. R., Fossett, T. R. D., Doyle, P. J., et al. (2007). Comparing connected language elicitation procedures in persons with aphasia: concurrent validation of the Story Retell Procedure. *Aphasiology* 21, 775–790. doi: 10.1080/02687030701189980

Mesulam, M. M. (2001). Primary progressive aphasia. *Ann. Neurol.* 49, 425–432. doi: 10.1002/ana.91

Newton, C., Acres, K., and Bruce, C. (2013). A comparison of computerized and paper-based language tests with adults with aphasia. *Am. J. Speech Lang. Pathol.* 22, 185–197. doi: 10.1044/1058-0360(2012/12-0027)

Nicholas, L. E., and Brookshire, R. H. (1993). A system of quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *J. Speech Hear. Res.* 36, 338–350. doi: 10.1044/jshr.3602.338

Noyes, J. M., and Garland, K. J. (2008). Computer- vs. paper-based tasks: are they equivalent? *Ergonomics.* 51, 1352–1375. doi: 10.1080/00140130802170387

Obler, L. K., and Gjerlow, K. (1999). *Language and the Brain.* Cambridge: Cambridge University Press.

Olness, G. S. (2006). Genre, verb, and coherence in picture-elicited discourse of adults with aphasia. *Aphasiology* 20, 175–187. doi: 10.1080/02687030500472710

Olness, G. S., and Englebretson, E. F. (2011). On the coherence of information highlighted by narrators with aphasia. *Aphasiology* 25, 713–726. doi: 10.1080/02687038.2010.537346

Olness, G. S., Matteson, S. E., and Stewart, C. T. (2010). "Let me tell you the point": how speakers with aphasia assign prominence to information in narratives. *Aphasiology* 24, 697–708. doi: 10.1080/02687030903438524

Olness, G. S., and Ulatowska, H. K. (2011). Personal narratives in aphasia: coherence in the context of use. *Aphasiology* 25, 1393–1413. doi: 10.1080/02687038.2011.599365

Palsbo, S. E. (2007). Equivalence of functional communication assessment in speech pathology using videoconferencing. *J. Telemed. Telecare* 13, 40–43. doi: 10.1258/135763307779701121

Papageorgiou, S. G., Economou, A., and Routsis, C. (2014). The 5 objects test: a novel, minimal-language, memory screening test. *J. Neurol.* 261, 422–431. doi: 10.1007/s00415-013-7219-1

Pritchard, M., Hilari, K., Cocks, N., and Dipper, L. (2018). Psychometric properties of discourse measures in aphasia: acceptability, reliability, and validity. *Int. J. Lang. Commun. Disord.* 53, 1078–1093. doi: 10.1111/1460-6984.12420

Prokopidis, P., and Piperidis, S. (2020). "A neural NLP toolkit for Greek," in *SETN 2020: 11th Hellenic Conference on Artificial Intelligence. September 2020* (Athens), 125–128.

Radanovic, M., Mansur, L. L., and Scaff, M. (2004). Normative data for the Brazilian population in the Boston diagnostic aphasia examination: Influence of schooling. *Braz. J. Med. Biol. Res.* 37, 1731–1738. doi: 10.1590/S0100-879X2004001100019

Saffran, E. M., Berndt, R. S., and Schwartz, M. F. (1989). The quantitative analysis of agrammatic production: procedure and data. *Brain Lang.* 37, 440–479. doi: 10.1016/0093-934X(89)90030-8

Stark, B. C. (2019). A comparison of three discourse elicitation methods in aphasia and age-matched adults: implications for language assessment and outcome. *Am. J. Speech Lang. Pathol.* 28, 1067–1083. doi: 10.1044/2019_AJSLP-18-0265

Stark, B. C., Dutta, M., Murray, L. L., Bryant, L., Fromm, D., MacWhinney, B., et al. (2021b). Standardizing assessment of spoken discourse in aphasia: a working group with deliverables. *Am. J. Speech Lang. Pathol.* 11, 491–502. doi: 10.1044/2020_AJSLP-19-00093

Stark, B. C., Dutta, M., Murray, L. L., Fromm, D., Bryant, L., Harmon, T. G., et al. (2021a). Spoken discourse assessment and analysis in aphasia: An international survey of current practices. *J. Speech Lang. Hear. Res.* 64, 4366–4389. doi: 10.1044/2021_JSLHR-20-00708

Stark, B. C., and Fukuyama, J. (2021). Leveraging big data to understand the interaction of task and language during monologic spoken discourse in speakers with and without aphasia. *Lang. Cognit. Neurosci.* 36, 562–585. doi: 10.1080/23273798.2020.1862258

Theodoros, D., Hill, A., Russell, T., Ward, E., and Wootton, R. (2008). Assessing acquired language disorders in adults via the Internet. *Telemed. J. e- Health* 14, 552–559. doi: 10.1089/tmj.2007.0091

UNESCO-UIS (2012). *International Standard Classification of Education-ISCED 2011*. UNESCO-UIS Institute for Statistics. Available online at: http://uis.unesco.org/sites/default/files/documents/international-standard-classification-of-education-isced-2011-en.pdf (accessed April 6, 2022).

Vajjala, S., and Meurers, D. (2012). "On improving the accuracy of readability classification using insights from second language acquisition," in *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, 163-173*. Montreal, QC: Association of Computational Linguistics. Available online at: https://aclanthology.org/W12-2019.pdf (accessed June 22, 2022).

Varlokosta, S., Stamouli, S., Karasimos, A., Markopoulos, G., Kakavoulia, M., Nerantzini, M., et al. (2016). "A greek corpus of aphasic discourse: collection, transcription, and annotation specifications," in *Proceedings of LREC 2016 Workshop. Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016), Monday 23rd of May 2016*, 14–21. Available online at: https://ep.liu.se/ecp/128/003/ecp16128003.pdf (accessed April 13, 2022).

Wallace, S. J., Worrall, L., Rose, T., Le Dorze, G., Breitenstein, C., Hilari, K., et al. (2019). A core outcome set for aphasia treatment research: The ROMA consensus statement. *Int. J. Stroke* 14, 180–185. doi: 10.1177/1747493018806200

World Health Organization (2001). *International Classification of Functioning, Stability and Health*. ICF. Available online at: https://apps.who.int/iris/handle/10665/42407 (accessed April 6, 2022).

Wright, H. H. (2011). Discourse in aphasia: an introduction to current research and future directions. *Aphasiology* 25, 1283–1285. doi: 10.1080/02687038.2011.613452

Wright, H. H., and Capilouto, G. J. (2009). Manipulating task instructions to change narrative discourse performance. *Aphasiology* 23, 1295–1308. doi: 10.1080/02687030902826844

frontiers | Frontiers in Psychology

TYPE Methods
PUBLISHED 20 April 2023
DOI 10.3389/fpsyg.2023.989499

OPEN ACCESS

EDITED BY
Natalia Gagarina,
Leibniz Center for General Linguistics (ZAS), Germany

REVIEWED BY
Kenneth Vaden,
Medical University of South Carolina, United States
Sarah Pila,
Northwestern University, United States

*CORRESPONDENCE
Rebecca Bright
✉ rbright@therapy-box.co.uk

SPECIALTY SECTION
This article was submitted to Psychology of Language, a section of the journal Frontiers in Psychology

RECEIVED 08 July 2022
ACCEPTED 31 January 2023
PUBLISHED 20 April 2023

CITATION
Bright R, Ashton E, Mckean C and Wren Y (2023) The development of a digital story-retell elicitation and analysis tool through citizen science data collection, software development and machine learning. *Front. Psychol.* 14:989499. doi: 10.3389/fpsyg.2023.989499

The development of a digital story-retell elicitation and analysis tool through citizen science data collection, software development and machine learning

Rebecca Bright[1]*, Elaine Ashton[2], Cristina Mckean[2] and Yvonne Wren[3,4,5]

[1]Therapy Box, London, United Kingdom, [2]School of Education, Communication and Language Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom, [3]North Bristol NHS Trust, Bristol, United Kingdom, [4]Bristol Dental School, University of Bristol, Bristol, United Kingdom, [5]Cardiff School of Sport and Health Sciences, Cardiff Metropolitan University, Cardiff, United Kingdom

Background: In order to leverage the potential benefits of technology to speech and language therapy language assessment processes, large samples of naturalistic language data must be collected and analysed. These samples enable the development and testing of novel software applications with data relevant to their intended clinical application. However, the collection and analysis of such data can be costly and time-consuming. This paper describes the development of a novel application designed to elicit and analyse young children's story retell narratives to provide metrics regarding the child's use of grammatical structures (micro-structure) and story grammar (macro-structure elements). Key aspects for development were (1) methods to collect story retells, ensure accurate transcription and segmentation of utterances; (2) testing the reliability of the application to analyse micro-structure elements in children's story retells and (3) development of an algorithm to analyse narrative macro-structure elements.

Methods: A co-design process was used to design an app which would be used to gather story retell samples from children using mobile technology. A citizen science approach using mainstream marketing *via* online channels, the media and billboard ads was used to encourage participation from children across the United Kingdom. A stratified sampling framework was used to ensure a representative sample was obtained across age, gender and five bands of socio-economic disadvantage using partial postcodes and the relevant indices of deprivation. Trained Research Associates (RA) completed transcription and micro and macro-structure analysis of the language samples. Methods to improve transcriptions produced by automated speech recognition were developed to enable reliable analysis. RA micro-structure analyses were compared to those generated by the digital application to test its reliability using intra-class correlation (ICC). RA macro-structure analyses were used to train an algorithm to produce macro-structure metrics. Finally, results from the macro-structure algorithm were compared against a subset of RA macro-structure analyses not used in training to test its reliability using ICC.

Results: A total of 4,517 profiles were made in the app used in data collection and from these participants a final set of 599 were drawn which fulfilled the stratified sampling criteria. The story retells ranged from 35.66s to 251.4s in length and had word counts ranging from 37 to 496, with a mean of 148.29 words. ICC between

the RA and application micro-structure analyses ranged from 0.213 to 1.0 with 41 out of a total of 44 comparisons reaching 'good' (0.70–0.90) or 'excellent' (>0.90) levels of reliability. ICC between the RA and application macro-structure features were completed for 85 samples not used in training the algorithm. ICC ranged from 0.5577 to 0.939 with 5 out of 7 metrics being 'good' or better.

Conclusion: Work to date has demonstrated the potential of semi-automated transcription and linguistic analyses to provide reliable, detailed and informative narrative language analysis for young children and for the use of citizen science based approaches using mobile technologies to collect representative and informative research data. Clinical evaluation of this new app is ongoing, so we do not yet have data documenting its developmental or clinical sensitivity and specificity.

KEYWORDS

story retell, citizen science, language sample, machine learning, speech pathology, story grammar

Introduction

The study of children's language acquisition has a long history. This fundamental developmental achievement has been scrutinised by scholars from many different disciplines, including psychology, linguistics, education and speech and language pathology. A key method to examine children's language learning, which has yielded crucial insights since the inception of audio-recording technology, is to record a sample of a child's interaction, transcribe the language heard, and analyse the linguistic structures used by the child (Bernstein Ratner and MacWhinney, 2019). Although not without challenges, recording and analysing a language sample has long been recognised as an ecologically valid measure of a child's language abilities in a functional context (Miller, 1996). In addition to research contexts, the analysis of language samples yields important insights for speech and language therapists/pathologists who work with individuals with language disorders.

Language samples and their place in clinical practice

Transcribing and analysing samples of a child's spoken language supports clinicians in evaluating performance with reference to typical development, undertaking goal-setting, and measuring progress. It is seen by some as the "gold standard" for analysing a child's language skills with advantages over standardised testing procedures, including a more naturalistic assessment of a child's ability and potentially providing less culturally biased measures of a child's development (Heilmann et al., 2010). It is also possible to repeat a language sample assessment more frequently than a standardised test without any threat to the validity or reliability of the procedure, and so enable evaluation of progress over time (Wilder and Redmond, 2022). Samples of narratives and story re-tells are particularly informative contexts for linguistic analysis in terms of their ability to distinguish between diagnostic subgroups (Botting, 2002) due to the high processing demands they place on the speaker to uncover impairments (Wagner et al., 2000). Importantly, they are also a very sensitive predictor of prognosis in both language and literacy outcomes in children with early language difficulties (Bishop and Edmundson, 1987; Botting, 2002; Miller et al., 2006). Analysis of narratives and story retells can focus on micro-structure elements, such as grammatical morphology, syntax and vocabulary and macro-structure features, related to the overarching organisation and coherence of the story, sometimes referred to as 'story grammar' (Westerveld and Gillon, 2010; Gillam et al., 2017). The former is highly informative to the clinician concerning the presence and nature of semantic and morpho-syntactic deficits and their impacts on functional communication; the latter brings insights related to discourse and pragmatic abilities.

Recent changes to diagnostic criteria for Developmental Language Disorder (DLD) bring a renewed focus on methods to evaluate a child's ability to use language *functionally* in context. A DLD diagnosis is not determined by cut-points on standardised tests but rather by a language problem that 'causes functional impairment in everyday life', (Bishop et al., 2017 p. 1068). Few rigorous and reliable assessment methods exist for identifying such functional impairments. Language sampling and analysis offer such a method; however, many barriers prevent its widespread use in clinical practice.

Barriers to the use of language sampling in practice

Despite numerous calls for clinical practice to change, so that language sampling, transcription and detailed analysis become standard practice, barriers of time, skills, knowledge and confidence levels continue to prevent this (Kemp and Klee, 1997; Westerveld, 2014; Pavelko et al., 2016; Pezold et al., 2020; Klatte et al., 2022). Training alone has been insufficient in leading to increased use of language sample analysis, despite clinicians having an awareness of the benefits (Klatte et al., 2022). Therefore, barriers other than skills and knowledge also need to be addressed.

The use of technology to support the use of language sample analysis in clinical practice

Computer-based language sample analysis is a way to gather qualitative information about a child's language that complements other assessment processes (Pezold et al., 2020; Klatte et al., 2022). Furthermore, the use of technology to semi-automate processes of transcription and analysis has the potential to ameliorate barriers of time and perhaps to scaffold and support clinicians who are less confident in linguistic analysis. However, despite the presence of existing software and training programs, clinicians report that the hurdles described above persist (Klatte et al., 2022). Hence, currently available tools are not yet suited to clinical practice in terms of ease of use and time demands. Klatte et al. (2022) suggested that language sampling software developed in codesign with clinicians and shorter narrative-based sampling could overcome some of the identified obstacles.

Challenges also exist in developing automated language analysis technology which can provide clinicians with the relevant analysis of micro and macro structures required to inform diagnosis and intervention. Micro-structure elements vary in the degree of challenge they present to automated analysis depending on the ambiguity and potential for miscategorisation. Identifying a determiner such as 'the' is relatively easy, a bound morpheme such as -ed is more complex, and a copula, whose identification rests on the surrounding context, is substantially more challenging. Macro-structure, or 'story grammar', is evaluated through the identification of the presence of the description by the speaker of factors such as the story setting, the initiating event, and the characters' internal response, together with a rating of the success or sophistication of the language used to describe those elements (Westerveld and Gillon, 2010). Potential automation to assist in this process requires the software to recognise the many different ways a speaker might encode an internal response or a setting and ascribe a relatively subjective rating to them.

The use of citizen science to support the development of an automated language analysis tool

Citizen science approaches involve members of the public as collaborators in scientific research, such as in formulating research questions, data collection or analysis of findings (Bonney et al., 2009). The relatively low cost of mobile app based data collection, high-quality audio recording, and attractive 'gamified' data elicitation procedures (Gillan and Rutledge, 2021) bring unprecedented opportunities to gather large-scale naturalistic language data. Furthermore, targeted marketing campaigns can enable geographical and socio-economic reach that may otherwise be difficult or costly. In this way, citizen science approaches enable the development and testing of novel software applications with large-scale data relevant to their intended clinical and research application. While citizen science approaches offer an attractive means of gathering data at low cost and quickly from a broad group of participants, limitations include variability in data and potential differences in how similar data would be collected in person by researchers. Here we examine the potential of such approaches to be used to develop a language sampling and analysis tool.

The current study

A product or tool that supports story retell elicitation, automated speech recognition, transcription improvement and language analysis is yet to be realised (Scott et al., 2022). This study takes the first steps in developing such a tool for language sample elicitation, collecting a large representative sample of young children's naturalistic language and developing and testing the app's ability to accurately analyse key aspects of the child's linguistic development.

The Language Explorer data collection app aimed to elicit a language sample *via* a story retell task and provide users with software-based tools to support transcription and analysis of micro and macro-structure elements of the samples. Supported by funding from an NIHR i4i Product Development Award, software was co-designed with children and clinicians. To develop a reliable and valid tool, we needed to collect large-scale data representing the likely range of ages and language abilities we would see in the clinical context for which the tool was intended. This would allow micro-structure analytical methods to be refined and a macro-structure analysis algorithm to be trained. To ensure Language Explorer could be used reliably in practice, we also needed to ensure it was acceptable to families. In 2020 we embarked on a Citizen Science study with two stages: (1) to collect a representative sample of United Kingdom children's story retelling using the Language Explorer app and (2) to complete the development of the language transcription and analysis tool.

We aimed to address the following research questions:

- Is it possible to gather a representative stratified sample of story retell recordings of children aged 4–7 years across the United Kingdom using Citizen Science methods?
- How acceptable is the Language Explorer App to families participating in the citizen science project?
- Is the quality of the recordings sufficient for reliable transcription and analysis?
- What level of reliability in automated micro-structure analyses can be achieved?
- Is it possible to develop a software platform that can provide reliable macro-structure analyses? If so, what level of reliability can be achieved?

The following presents the methods and results for each stage of the citizen science project, including data collection and analysis. The clinical evaluation of the tool will be reported in later publications.

Methods

The study had four phases (1) design and development; (2) language sample data collection; (3) data analysis and software refinement and (4) software testing. Phases 3 and 4 involved different methods for the macro and micro-structure elements. The phases, their linkage and their sub-components are represented in Figure 1.

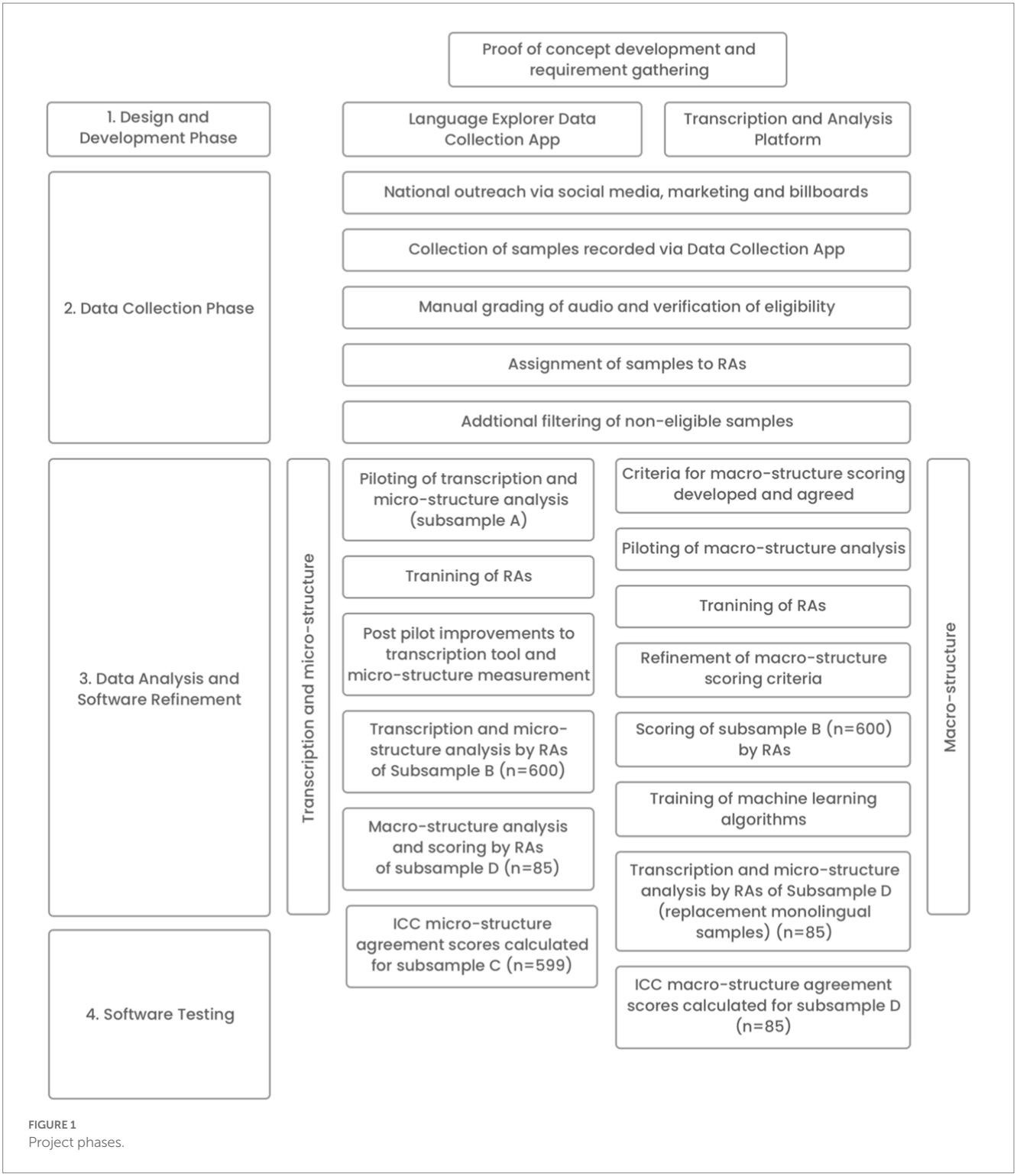

FIGURE 1
Project phases.

Design and development

Codesign of the language explorer data collection app

Using principles of user-centred design and co-design, the design team worked with clinicians, parents and children of primary school age to develop the content for the story retelling stimulus. The semi-animated story of a boy on a treasure quest underwent usability testing with children. Using a standard usability testing approach (Gomoll, 1990; Norman and Kirakowski, 2018) children, clinicians and parents at two Hackney schools were provided with the app on iPads, given an overview of the app's purpose, provided with instructions to use the app from start to finish and observed using it. Verbal feedback and observations relating to engagement, accessibility and ease of use were collected. Based on the usability testing, instructions were refined, button

sizes adjusted and user experience design elements were added to make progressing through the app more intuitive. The story script was developed following advice from researchers with specialist knowledge of syntactic structures likely to be challenging to children with language disorders. A survey to elicit parent feedback during the citizen science phase was also included (see Appendix 1).

Codesign of the transcription and micro-structure analysis platform

Proof of concept work examined the potential for using speech recognition for child language sampling and analysis was completed. 'Requirement gathering' was completed to determine what a project needs to achieve and what needs to be created to make that happen. Feedback was elicited from clinicians on early-stage prototypes with iterative improvements in the design of the transcription improvement tool (to manually improve the accuracy of transcription provided by the Automatic Speech Recognition software) and micro-structure analysis software between workshops.

Language sample data collection

A United Kingdom wide campaign was undertaken to call for participants to crowdsource samples using the Language Explorer app. Ethical approval was provided by Bristol University (reference 97,304). Outreach *via* social media, press and paid targeted marketing was conducted. In addition, the use of location-targeted billboards was designed to attract attention to the study. The outdoor media was placed to take advantage of the expected traffic of parents with children within the target age range, including within a short range of schools and transport hubs. Campaign messaging encouraged participants to contribute to the study to help children with language disorders in the future.

The app was downloadable from the AppStore and PlayStore. Consent for the data to be used for research was sought *via* the app. Recordings were transferred to a designated data management platform meeting GDPR requirements. Families could complete the task offline with data only uploaded when they were next online to reduce reliance on connectivity. In addition to the audio recordings, demographic data were entered in the app by the end-user, presumed to be the parent/carer key to enable stratification of the sample and consideration of exclusion and inclusion criteria for data analysis. These were the child's age, country, partial postcode, whether the child had a diagnosed communication difficulty or disorder or other disability and the languages spoken in the home (see Appendix 1). A proportionate stratified sampling approach was used to gather participants to use in the piloting and training of the software such that it would cover children across the United Kingdom equally distributed by sex, five age bands (4:0–4:5; 4:6–4:11; 5:0–5:11; 6:0–6:11; and 7:0–7:11) and quintiles of socio-economic disadvantage (see Table 1 for planned sample). The latter was defined using partial postcodes. Partial postcodes were used to enable stratification whilst remaining at a level of granularity unlikely to raise concerns amongst participants regarding confidentiality and data protection. The 2019 Indices of Multiple Deprivation (IMD) for each United Kingdom nation were consulted (McLennan et al., 2019). National quintiles for each partial postcode area were created by averaging the IMD ranking of all postcodes represented within the partial postcode area. Participants' partial postcodes were then mapped to these quintiles.

Samples from outside of the United Kingdom or where children spoke a language other than English were excluded from further consideration at this phase, as were children with an identified disability or communication disorder. The focus on monolingual typically developing children at this stage was to enable valid comparison of these data to the planned clinical evaluation population. Recruitment continued until all strata contained the target numbers with the desired characteristics. For example, 20 children aged 4–4:05 in each IMD quintile made up of 10 girls and 10 boys (see Table 1).

Sample achieved

In achieving the stratified sample, 4,517 profiles were made in the Language Explorer data collection app (Figure 2). Following the exclusion of 329 profiles registered as being from outside Great Britain or Northern Ireland, a total of 4,188 profiles remained. Of the 4,188 profiles, 2,340 had completed the story retelling, sentence comprehension and repetition tasks. The initial exclusion of participants who were not English first language participants and/or listed as having a communication or other disability left 1,451 samples. A further 312 participants that would have otherwise been eligible were excluded as they were not graded as audible when samples were screened manually. Samples graded as audible totalled 889.

The obtained recordings were used for differing purposes with different subsamples of children drawn from the pool of 889 as appropriate to the purpose of the work.

TABLE 1 Target stratified sampling frame.

Age	4–4:05	4:06–4:11	5:0–5:05	5:06–5:11	6–6:11	7–8
	$N = 100$	$N = 100$	$N = 100$	$N = 100$	$N = 100$	$N = 100$
IMD Q1	20	20	20	20	20	20
IMD Q2	20	20	20	20	20	20
IMD Q3	20	20	33	20	20	20
IMD Q4	20	20	20	20	20	20
IMD Q5	20	20	20	20	20	20
	M (50)	M (50)	M (50)	M (50)	M (50)	M (50)
	F (50)	F (50)	F (50)	F (50)	F (50)	F (50)

IMD, indices of multiple deprivation (Office for National Statistics, 2019); *Q*, quintile where 1 is the least deprived and 5 is the most deprived.

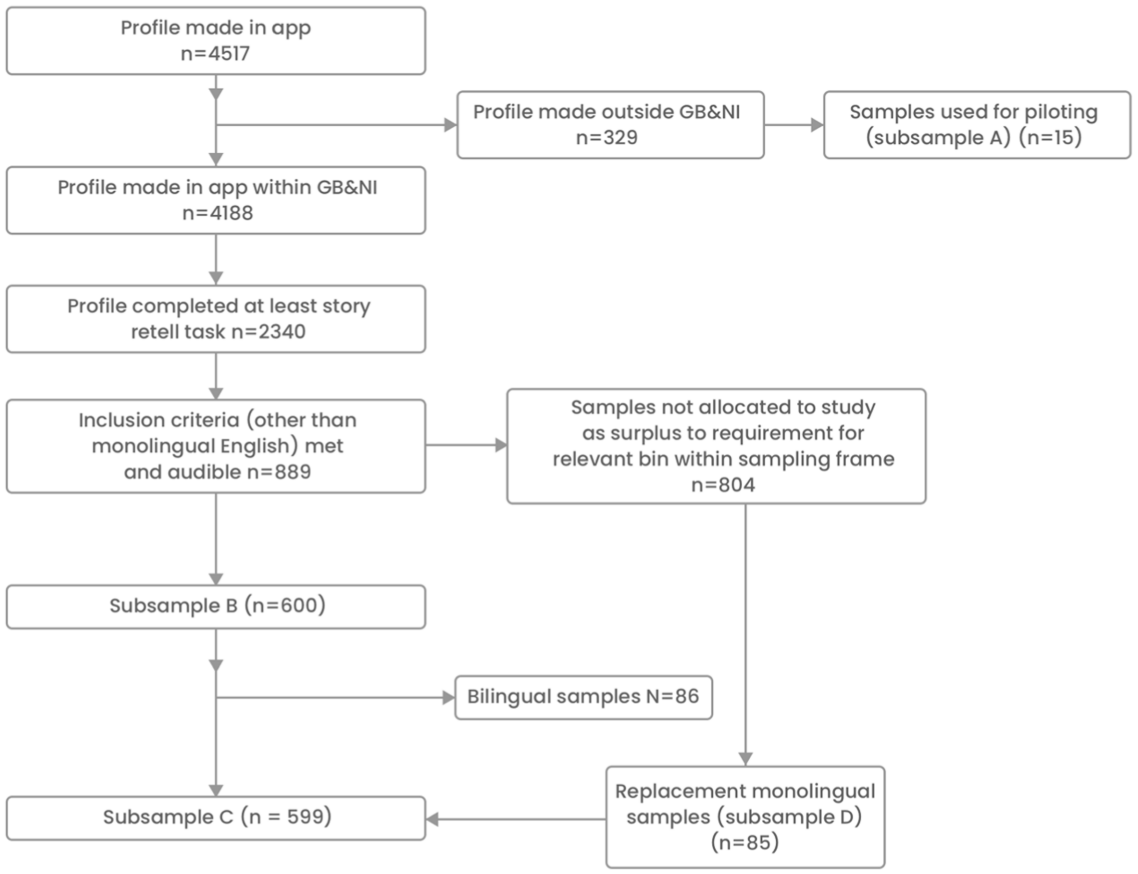

FIGURE 2
Participant flow chart.

Subsample A: Piloting and micro-structure analysis refinement – a sample of 15 English-speaking children from outside of the United Kingdom surplus to requirement for the stratified sample but of sufficient quality for piloting and for RA training.

Subsample B: Development of the macro-structure algorithm – 600 children either monolingual or bilingual with English as a first language who met our stratification strategy with respect to gender, IMD and age. This was made up of 86 bilingual children and 514 monolingual children. It was proving difficult to identify monolingual children fitting precisely into the required stratified sampling frame with respect to age, gender and IMD. Including English first language bilingual children allowed the development of the algorithm to proceed whilst monolingual replacement samples were sought.

Subsample C: Testing of the micro-structure analysis – 599 monolingual children who met our stratification strategy with respect to age, IMD and gender. A monolingual only sample was required for the testing phase to better align with the participants expected to be recruited for the later clinical evaluation. Therefore this sample is made up of the 514 monolingual children in subsample B plus new monolingual children with the same age, gender, IMD characteristics as the 86 bilingual children dropped from subsample B. We identified 85 appropriate monolingual children giving a total of 599 in this subsample.

Subsample D: Testing of the macro-structure analysis – 85 monolingual children not used in the development of the macro-structure analysis algorithm also used in subsample C.

Data analysis and software refinement – Micro-structure

Anonymous recordings were received by the research team in the data management platform. Data were then analysed by a tea of four research associates: three junior research assistants (RAs) who were all clinically qualified SLTs and one lead RA, also a qualified SLT and with an additional qualification in linguistics, all based at Newcastle University. Following checking of audio, samples in each stratum were allocated to one of the three RAs, each completing transcription and analysis. RA1 completed transcribed and analysed 36% of the samples, RA2 44% ad RA 3 19%. The lead RA carried out reliability checking of 10% of samples. RAs were 'blind' to the age, gender and other demographic data of the sample. A further quality assurance process took place at this stage, where the RAs judged several samples not to be suitable for language analysis. Samples were excluded where the story's content was too short, the narrative was incomplete, where it was judged there was too much secondary speaker input, which limited the child's performance, for example, the adult retold the story

and the child repeated what the adult had said, or where a non-English language was spoken. These samples were flagged and removed, and replacement samples meeting the same demographic criteria necessary for the stratification sample were allocated. The RAs discussed with the lead RA if they had questions about the inclusion of a sample.

The goal was to compare the Language Explorer analysis of micro-structure components to the Systematic Analysis of Language Transcripts (SALT) software (Miller et al., 2019). SALT is the most widely used language analysis software designed for clinical use and has high levels of validity and reliability (Tucci et al., 2022). The SALT software requires the researcher or clinician to transcribe and annotate the transcript following particular conventions and then automatically calculates certain micro-structure features in language samples, e.g., the number of adjectives and prepositions. It also allows the user to manually mark other morphological markers that it can then calculate automatically, e.g., plural -s, past tense -ed, and personalised tags such as auxiliary and copula verbs. The RAs needed to be both reliable and consistent in their transcription and utterance segmentation and in the conventions required by the SALT software for accurate analysis. These consist primarily of additional 'tags' required to identify key micro-structures in the sample. The three RAs were trained using subsample A - pilot dataset of 15 samples which were not included in any further analysis. The aim was to achieve greater than 85% inter-rater agreement in manual transcription and SALT language analysis before moving on to the samples to be used in the later phases of the method. The lead RA compared each of the RAs' transcription and analyses of the pilot data with her own, and they reached a level of 97% agreement for the manual transcription and 96% for the micro-structure language analysis completed in SALT, indicating reliable use of transcription conventions across the team.

Using the same subsample A ($N = 15$), the lead RA compared the transcripts from the manually completed pilot samples and the micro-structure analysis from SALT with the samples completed using the Language Explorer clinical software tools. This involved a comparison of transcription using the transcription improvement tool and reviewing the counts of the 'parts of speech'. Feedback was provided to the software engineers, and improvements were made to the software. Automated measures that did not reach the 85% level of agreement between manual and automated results were scrutinised, and the potential sources of error were discussed to retrain the automated microstructural analysis using the software. The software engineers and the lead RA checked each problematic metric in the 15 pilot samples. The software was modified as a result, including clarifying rules in the software for identifying metrics and providing examples.

Data analysis and software development: Macro-structure

Unlike the micro-structure analysis, the macro-structure analysis required ongoing refinement using subsample A and subsample B. The macro-structure metrics focussed on seven macro-structure 'story grammar' elements (setting, initiating event, internal response, plan, attempt, consequence and character) based on Stein and Glenn (1979). A matrix of definitions and examples was prepared using a scoring system of 0–3 for each element, with 0 being unobserved and 3 being the score allocated for a full demonstration of that macro-structure element. This was used to build and train the algorithms.

Before training the algorithms, high levels of agreement between the RAs were necessary to ensure high-quality data. Substantial training and refinement of the scoring rubric were required to reach the necessary levels of agreement between the RAs. Initially, the levels of agreement of the story grammar scoring between the lead RA and the three RAs were low (ICC of RA1 0.455, RA2 0.562, RA3 0.587). The lead RA therefore refined the macro-structure descriptors and scoring examples using specific examples from the pilot dataset stories, as well as applying learning from other studies (Beswick, 2008; Gillam et al., 2017; Gillam and Gillam, 2018; Jones et al., 2019; Diehm et al., 2020). Due to the more subjective nature of macro-structure scoring and hence challenges in establishing reliability (Calder et al., 2018), a threshold was set for inter-rater agreement in the training phase of 75% for each story grammar component and 85% agreement for the story grammar total score Once this was achieved, the RAs could move on to scoring subsample B ($N = 600$) to train the algorithm.

The training of the RAs used real examples from the pilot to further support learning. It worked in short intervals using sets of three samples from subsample A ($N = 15$) before checking in on agreement and discussing sources of disagreement. A final test set of five additional pilot samples was used following this revised training. The two RAs achieved above the necessary agreement scores with the lead RA (0.885 and 0.940 for the macro-structure (story grammar) elements and 0.938 and 0.938 for the total macro-structure score), noting that one RA left the project at this point on maternity leave. The RAs then scored subsample B ($N = 600$) for story grammar, with the lead RA providing reliability checking of 10% of the samples. The agreement scores for this reliability checking were 0.907 and 0.869 for each of the two RAs for story grammar components and 0.956 and 0.925 for the total story grammar score (Appendix 5). Following completion of the manual scoring, the revised descriptors and scoring data for each macro-structure element and the examples were used to train the algorithms.

Software testing: Micro-structure

The RAs completed orthographic transcription and manual analysis of the samples received in the stratified subsample B ($N = 600$). Further reliability checking was carried out with 10% of the transcriptions and SALT language analyses for each RA compared with the lead RA using an identical method to that used for piloting (60 samples in total). The levels of reliability achieved were 93% for the transcription and 98% for the language analyses (Appendix 5), confirming the maintenance of the high levels of reliability achieved during training. Subsample C ($N = 599$) using only monolingual participant samples was used for the analysis of the reliability of the Language Explorer tool for micro-structure features. Intra-class correlations (ICCs) were calculated by comparing each of the parts of speech metrics calculated by the Language Explorer tool and those calculated for comparison *via* manual transcription and SALT analysis.

Software testing: Macro-structure

Given the need to use the whole of subsample B ($N = 600$) in the training of the algorithms, testing the performance of the macro-structure analysis module in the clinical tools software was limited to

FIGURE 3
Screen from the language explorer story in the app.

using subsample D ($N = 85$ monolingual replacement samples) that were not used in training. Intra-class coefficient (ICC) scores were calculated for each macro-structure element and the total score for the seven elements.

Results

Results are presented in turn for phases (1) design and development; (2) language sample data collection, and (3) software testing.

Design and development

The co-design work with children, parents, researchers and clinicians informed the development of the final Language Explorer Mobile application. This app presents semi-animated story of a boy on a treasure quest and following the story the child is asked to retell the story with pictorial support. This story retell is recorded using inbuild recording technology in the phone or tablet being used. The resulting story was designed to elicit a linguistically rich sample with maximum efficiency. Participants' parents using the app in the citizen science phase were surveyed and asked about their experience. A total of 426 parents completed the survey. Most parents reported that the app was easy to use (95%) and that their children enjoyed using it (91%) (Figure 3).

The workshops with clinicians provided insights into the desire for technology to make language sampling quicker and easier and the need to collect samples on contemporary mobile technologies. The preference was for any such technology not to be a 'black-box' system that produces a 'score' without a transparent method but for clinicians to review the analysis and understand the processes and metrics. That is for the app to provide familiar and readily interpretable micro- and macro-structure metrics.

With regards to transcription processes the following features were developed as a result of the co-design and consultation. Upon receiving an audio recording of the story retell from the app, an automated speech recognition (ASR) based transcript is created and presented to the clinician user. Given the current industry accuracy for ASR for child speech (Yeung and Alwan, 2018; Hair et al., 2019), there remains a need to correct and improve the transcriptions. A 'transcription improvement tool' was designed and built to allow clinicians to use the ASR transcription and select words or utterances transcribed accurately by the ASR while making any corrections. The clinician at this stage also follows specific conventions to segment utterances, identify secondary speaker utterances and annotate occurrences of mazes, mispronunciations or unintelligible speech. These are necessary to ensure the analysis software can produce accurate micro-structure metrics consistently across speakers and across users checking the transcriptions. After confirming an accurate transcript, clinicians are presented with the same transcript. It is colour-coded by parts of speech with the marking of morpheme boundaries, allowing additional parts of speech to be tagged. There is the possibility at this stage of checking and modification by the clinician (Figure 4).

Data collection

The geographical distribution across United Kingdom regions of Subsample C used for testing the reliability of the micro-structure analysis was highly similar to that of the United Kingdom population across those regions (Table 2).

Characteristics of subsample C ($N = 599$) used to test the micro-structure analysis are described in Table 3 Speech duration ranged from 35.66 s to 251.4 s. Table 3 summarises the duration of the samples, their length in total number of words and the number of secondary speaker utterances.

Software testing: Micro-structure

The following presents the ICC scores for each micro-structure analysis metric between the RAs and the Language Explorer output computed for subsample B ($N = 599$) (Table 4). Note the use of full

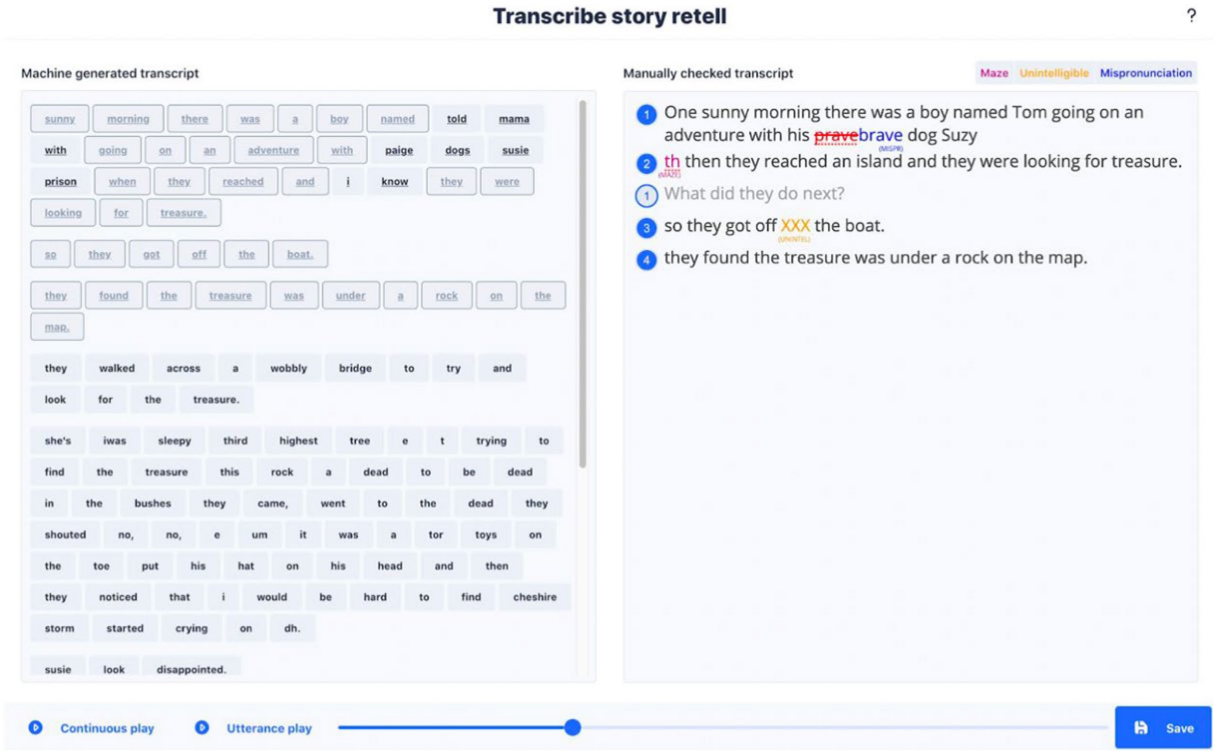

FIGURE 4
Screen from the transcription improvement tool.

TABLE 2 Participant geographical location based on partial postcode compared with population distribution.

United Kingdom region	Participants	Language explorer sample representation	Population distribution	Difference between proportion in sample obtained and United Kingdom population
East Midlands	53	8.8%	7.23%	1.6%
East of England	43	7.2%	9.25%	−2.1%
London	62	10.4%	12.97%	−2.6%
North East	42	7.0%	4.12%	2.9%
North West	78	13.0%	11.14%	1.9%
Northern Ireland	8	1.3%	2.83%	−1.5%
Scotland	30	5.0%	8.38%	−3.4%
South East	67	11.2%	13.63%	−2.4%
South West	82	13.7%	8.37%	5.3%
Wales	20	3.3%	4.82%	−1.5%
West Midlands	57	9.5%	8.91%	0.6%
Yorkshire and The Humber	57	9.5%	8.36%	1.2%
Total	599			

transcript and analysis set scoring differs in that the analysis set scores were calculated after the removal of incomplete utterances and utterances containing unintelligible segments or mazes (Table 4). The analysis set is the most valid measure of a child's language skills compared to the full analysis set, which may suggest the child can produce longer utterances than they can with the inclusion of mazes, etc.

From a total of 44 metrics, the only ICC scores that fell below the 'good' level of reliability (<0.70) were relative pronouns, present progressives and irregular plurals.

TABLE 3 Duration of speech, length in number of words and number of secondary speaker turns by age bands in the "Final 599" sample.

Age band	Length in speech duration in seconds M (SD)	Length in total number of words M (SD)	Number of Secondary speaker utterances M (SD)
	N = 588	*N* = 599	*N* = 599
4:0–4:5	109.48 (37.1)	136.18 (48.9)	11.36 (15.1)
4:6–4:11	106.13 (35.7)	132.58 (45.6)	8.75 (10.4)
5:0–5:5	108.62 (33.7)	145.71 (43.9)	7.26 (10.4)
5:6–5:11	108.53 (34.3)	146.19 (42.9)	4.54 (7.1)
6:0–6:11	109.38 (34.2)	162.67 (54.4)	2.81 (4.3)
7:0–7:11	107.47 (32.1)	167.54 (40.8)	2.42 (4.2)

TABLE 4 Micro-structure metric ICC scores.

Metric	ICC	Metric	ICC
Mean length of utterance (words) full transcript	0.911	Present progressives	0.434
Mean length of utterance (morphemes) full transcript	0.832	Questions	0.986
Max length of utterance (words) full transcript	0.986	Subordinate conjunctions	0.772
Max length of utterance (morphemes) full transcript	0.954	Coordinating conjunctions	1.000
Mean length of utterance (words) analysis set	0.904	Regular -s plurals	0.995
Mean length of utterance (morphemes) analysis set	0.924	Irregular plurals	0.664
Max length of utterance (words) analysis set	0.755	`s possessive	0.843
Max length of utterance (morphemes) analysis set	0.793	Articles	0.999
Total utterances	1.000	Regular past tense (−ed)	0.968
Total words	0.998	Irregular past tense	0.984
Keywords	0.987	Third person regular, present tense	0.934
Synonyms of keywords	0.912	Third person irregular, present tense	0.986
Type-token ratio	0.955	Unintelligible words	0.999
Nouns	0.975	Intelligibility	0.998
Pronouns	0.998	Count of mazes	1.000
Verbs	0.987	Comparatives	0.973
Relative pronouns	0.213	Superlatives	0.990
Adverbs	0.884	Contractible copula	0.957
Adjectives	0.949	Uncontractible copula	0.955
Determiners	0.970	Contractible auxiliary	0.904
Particles	0.850	Uncontractible auxiliary	0.870
Prepositions	0.965	Words per minute (*N* = 588)	0.927

ICC reliability = <0.50 = poor; 0.50–0.70 = moderate; 0.70–0.90 = good; >0.90 = excellent (Koo and Li, 2016).

Software testing – Macro-structure analysis module

ICC scores were calculated (Table 5) using subsample D (*N* = 85).

The setting, initiating event, plan and consequence elements had ICC scores categorised as good. Internal response and attempt had ICC scores classified as moderate reliability, and Character had a reliability level of >0.90, which is designated as excellent. When compiled as a total score as a composite of each of the seven macro-structure elements, the total macro-structure story grammar score had an ICC of 0.928, which is classed as excellent reliability (Koo and Li, 2016).

Discussion

The study demonstrates that it is possible to gather a representative stratified sample of story recall recordings of children aged 4–7 years in the United Kingdom using mobile technology and Citizen Science methods with recordings of sufficient quality for reliable transcription and analysis. Our findings suggest that substantial oversampling is required for such methods to succeed. Approximately 66% of the participants who signed up from the United Kingdom completed all the necessary tasks in the Language Explorer App, including the Story Retell task. Of those 2,340 complete samples, 889 (38%) met the inclusion

TABLE 5 Macro-structure metrics ICC scores for monolingual replacement samples (n =85).

Macro-structure metric	ICC
Setting	0.850
Initiating event	0.898
Internal response	0.577
Plan	0.841
Attempt	0.616
Consequence	0.729
Character	0.939
Composite macro-structure "story grammar" score*	0.928

ICC reliability = <0.50 = poor; 0.50–0.70 = moderate; 0.70–0.90 = good; >0.90 = excellent (Koo and Li, 2016).

criteria and were recorded with sufficient quality to be audible. Hence in terms of usable recordings, researchers using these approaches would likely need to oversample by a factor of 2.6. The sample reduces still further when additional exclusion criteria are applied. However, given the low cost of this approach and speed of data acquisition, the method could be of substantial interest to the child language research community. These findings, therefore, address our first two research questions and provide additional information to guide future work of this kind.

Two key caveats must be considered when choosing this method and interpreting our data. First, the partial postcode approach means that the sample is likely to be slightly more advantaged than the United Kingdom population as a whole, as it is likely that more advantaged families within each partial postcode grouping would participate. However, they are likely significantly more representative than many studies in the field of child language, given the speed and low resources needed to recruit families in lower SES postcode areas when compared to the difficulties often experienced by researchers to reach these groups, our data suggest that using a Citizen Science approach using social media, press and paid targeted marketing approaches holds promise for the recruitment of families who are traditionally under-represented in research.

Second, despite instructions not to help children, parents scaffold their child's narratives to varying degrees to support them in completing the story retelling. Children learn the skill of creating and retelling narratives through social interaction and parents/caregivers' engagement in narrative co-construction with their child (Wood et al., 1976; Stein and Glenn, 1979), providing scaffolding to support the child to extend and increase the sophistication of their narratives over development. For example, a parent may prompt the child to produce the setting component of the narrative (e.g. 'where were they going?') (Stein and Glenn, 1979). This prompt is provided until the child internalises the skill and can use the setting component in their narrative without support. This scaffolding from parents naturally decreases over development in response to the child's increasing abilities (Bailey and Moughamian, 2007; Bailey et al., 2020). This variability in the implementation of a task is a risk in all Citizen Science data collection (Borda, 2019). A balance must be struck between the benefits of large-scale, low-cost data collection and some variability in task implementation. Our samples have high ecological validity regarding the nature of co-constructed narratives over this developmental period. However, this co-construction makes comparing samples elicited in a clinical context more challenging.

Turning to the research questions regarding the reliability of the automated analysis of micro and macro structure components of the story retell after automatic transcription has been checked and corrected. A set of agreed transcription conventions followed, and the reliability of the micro-structure metrics yielded from the Language Explorer software when compared to SALT software was mostly high. Of the 44 metrics, 33 were excellent, eight were good, one was moderate (irregular plurals), and two were poor (present progressive and relative pronouns). The total macro-structure score ICC, when compared to rating by a trained SLT, was also excellent, indicating it is possible to develop a software platform which can provide reliable macro-structure analyses for a specific story retell. Indeed, the software had good or excellent reliability for all macro-structure elements excepting 'internal response' and 'attempt' components, suggesting it could provide useful clinical information regarding the overall quality of a child's narrative macro-structure abilities. We, therefore, recommend further work with clinicians to decide whether some of the least reliable metrics could potentially be dropped entirely from the app's final reporting output if they are not particularly clinically informative. Also, the final clinical version of the app includes instructions for clinicians regarding the metrics that require manual checking and how to do that.

It must be noted, however, that we have not tested the reliability of these scores regarding the degree to which they represent the child's broader abilities. There is no explicit agreement in the literature regarding the length of a narrative story retell or spontaneous language samples which provide reliable estimates of a child's wider abilities (Heilmann et al., 2010; Guo and Eisenberg, 2015; Wilder and Redmond, 2022). In the present study, the narratives that had duration data (n = 588) ranged from 35.66 s to 4 min and 19 s, with a mean of 1 min and 48 s (108.26 s). Recently Wilder and Redmond (2022) demonstrated that several language sample metrics (including MLU and number of different words) reach acceptable levels using 3- and 7-min language samples when compared to metrics obtained in 20-min samples. Also, Heilmann and colleagues have demonstrated stable results for productivity and MLU from narrative and other samples of 1–3 min (Heilmann et al., 2010, 2013). Further work to test representativeness compared to a child's wider language use of the language elicited by narrative retells in general and Language Explorer, in particular, is warranted.

The reliability of the data provided by the Language Explorer App also rests on the accuracy with which the SLT or researcher checks and prepares the language transcript. Following the conventions for utterance segmentation, correctly marking unintelligible utterances, mazes etc., is essential for reliable metrics to be calculated (see Appendix 4). They will also need support to check those few metrics with low reliability identified above. Training materials regarding transcription and analysis checking will therefore need to be included to support clinicians in using the app reliably. Additional work to evaluate this training and other steps in the clinical application of Language Explorer is underway and will be reported elsewhere. Further work would also be needed to assess the clinical use of the story comprehension and repetition subtests.

In terms of acceptability, extremely high numbers of parents reported their children enjoyed the app and found it easy to use. This supports its potential success in its current form for research purposes and is promising in terms of its potential for application in clinical practice. However, Language Explorer will be implemented slightly differently in the clinical context by SLTs, and the acceptability and feasibility of its use in that context will be tested in the clinical evaluation study currently underway.

Strengths and limitations

The study recruited large numbers of children across a range of socio-economic quintiles and with a wide geographical spread. Furthermore, the majority (66%) of those who signed up completed the tasks within the App. As identified above, due to the use of partial rather than full postcodes, a bias towards more socially advantaged groups than the United Kingdom population is a possible issue (i.e., with the higher SES within each quintile possibly being recruited). However, compared to other research methods and study samples, the Citizen Science approach, linked with a targeted and multi-strategy marketing campaign, appears to be a cost-effective method for reaching subgroups often considered 'harder to reach' using more traditional recruitment methods.

Parental scaffolding of narrative retells creates issues comparing these data with retells elicited in more controlled clinical contexts. However, they represent an ecologically valid representation of co-constructed narratives, which form a crucial stage in typical narrative development.

Rigorous training of the RAs and high levels of transcription and analysis reliability among the researchers, together with the large sample (599) to test the micro-structure metrics' accuracy, provide significant confidence in the study findings. The macro-structure/story grammar analysis module also benefited from the quality and quantity of the data needed to inform the algorithm of the descriptors of each of the seven story grammar elements and the examples for each. A total of 600 samples were used to train the algorithm. Testing with only 85 samples that were not used in training is a limitation. However, the agreement scores for the story grammar elements are promising. In particular, the composite total macro-structure, story grammar score appears robust to measure a complex, discourse-level language feature. It has the potential to be developed to be used to guide assessment in clinical practice, functioning as an indicator of the need or otherwise to examine macro-structure abilities in more detail.

Conclusion

Language sampling analysis is considered best practice for speech and language therapy assessment of child language (Miller, 1996). The barriers of time and the need for intuitive software to make the process variable for clinicians are established (Klatte et al., 2022) Work presented here has demonstrated the potential of semi-automated transcription and automated linguistic analyses to provide reliable, detailed and informative narrative language analysis for young children and for the use of mobile technologies to collect representative and informative clinical and research data.

A feasibility study of Language Explorer modules is currently underway in clinical settings using the elicitation app and the transcription and analysis tools developed using this dataset. Guidance and training materials were also designed to enable clinical researchers to adhere to the transcription conventions required for reliable automated analyses to be completed in this evaluation phase. This evaluation in a clinical context aims to identify any potential benefits which Language Explorer can bring to clinical practice and what further work would be required to realise them. In this way, we aim to remove critical barriers to using narrative language sample analysis to practice and bring the benefits of more detailed, functional, ecologically valid and sensitive language assessment to the assessment and therapy planning for children with DLD.

Data availability statement

The datasets presented in this article are not readily available because participants did not provide consent for such access. Requests to access the datasets should be directed to, rbright@therapy-box.co.uk.

Ethics statement

The studies involving human participants were reviewed and approved by Faculty of Health Sciences Research Ethics Committee University of Bristol. Consent to participate in this study was provided by participants' parents/carers via the app.

Author contributions

RB, CM, YW, and EA contributed to the conception and design of the study. EA performed the statistical analysis. RB wrote the first draft of the manuscript. CM and EA wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study/project is funded by the NIHR i4i (NIHR200889). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Conflict of interest

RB is the co-founder of the commercial organisation, Therapy Box, which will commercialise the software described in the article.

The remaining 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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.989499/full#supplementary-material

References

Bailey, A. L., and Moughamian, A. C. (2007). Telling stories their way: narrative scaffolding with emergent readers and readers. *Narrat. Inq.* 17, 203–229. doi: 10.1075/ni.17.2.04bai

Bailey, A. L., Moughamian, A. C., Kelly, K. R., McCabe, A., and Huang, B. H. (2020). Leap-frog to literacy: maternal narrative supports differentially relate to child oral language and later reading outcomes. *Early Child Dev. Care* 190, 1136–1149. doi: 10.1080/03004430.2018.1521807

Bernstein Ratner, N., and MacWhinney, B. (2019). "TalkBank resources for psycholinguistic analysis and clinical practice," in *Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences*. eds. A. Pareja-Lora, M. Blume and B. Lust (MIT Press).

Beswick, N. (2008). Determination of the Inter-rater Reliability of the Edmonton Narrative Norms Instrument. *Department of Speech Pathology and Audiology Edmonton Alberta*.

Bishop, D. V. M., and Edmundson, A. (1987). Language-impaired 4-year-olds: distinguishing transient from persistent impairment. *J. Speech Hear. Disord.* 52, 156–173. doi: 10.1044/jshd.5202.156

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., and Greenhalgh, T.the CATALISE-2 Consortium (2017). Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: terminology. *J. Child Psychol. Psychiatry* 58, 1068–1080. doi: 10.1111/jcpp.12721

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., et al. (2009). Citizen science: a developing tool for expanding science knowledge and scientific literacy. *Bioscience* 59, 977–984. doi: 10.1525/bio.2009.59.11.9

Borda, A. (2019). Citizen science models in health research: an Australian commentary. *Online J. Public Health Inform.* 11:e23. doi: 10.5210/ojphi.v11i3.10358

Botting, N. (2002). Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Lang. Teach. Ther.* 18, 1–21. doi: 10.1191/0265659002ct224oa

Calder, S. D., Claessen, M., and Leitão, S. (2018). Combining implicit and explicit intervention approaches to target grammar in young children with Developmental Language Disorder. *Child Language Teaching and Therapy* 34, 171–189. doi: 10.1177/0265659017735392

Diehm, E. A., Wood, C., Puhlman, J., and Callendar, M. (2020). Young children's narrative retell in response to static and animated stories. *International Journal of Language & Communication Disorders* 55, 359–372. doi: 10.1111/1460-6984.12523

Gillam, R. B., and Gillam, S. L. (2018). *Assessment of Narratives in School-age Children*. Oregon, USA: OSHA presentation.

Gillam, S. L., Gillam, R. B., Fargo, J. D., Olszewski, A., and Segura, H. (2017). Monitoring indicators of scholarly language: a progress-monitoring instrument for measuring narrative discourse skills. *Commun. Disord. Q.* 38, 96–106. doi: 10.1177/1525740116651442

Gillan, C. M., and Rutledge, R. B. (2021). Smartphones and the neuroscience of mental health. *Annu. Rev. Neurosci.* 44, 129–151. doi: 10.1146/annurev-neuro-101220-014053

Gomoll, K. (1990). "Some techniques for observing users" in *The Art of Human-Computer Interface Design*. eds. B. Laurel and S. J. Mountford (Addison-Wesley), 85–90.

Guo, L.-Y., and Eisenberg, S. (2015). Sample length affects the reliability of language sample measures in 3-year-olds: evidence from parent-elicited conversational samples. *Lang. Speech Hear. Serv. Sch.* 46, 141–153. doi: 10.1044/2015_LSHSS-14-0052

Hair, A., Ballard, K. J., Ahmed, B., and Gutierrez-Osuna, R. (2019). Evaluating Automatic Speech Recognition for Child Speech Therapy Applications. The 21st International ACM SIGACCESS Conference on Computers and Accessibility, 578–580.

Heilmann, J., DeBrock, L., and Riley-Tillman, T. C. (2013). Stability of measures from Children's interviews: the effects of time, sample length, and topic. *Am. J. Speech Lang. Pathol.* 22, 463–475. doi: 10.1044/1058-0360(2012/11-0035)

Heilmann, J., Nockerts, A., and Miller, J. F. (2010). Language sampling: does the length of the transcript matter? *Lang. Speech Hear. Serv. Sch.* 41, 393–404. doi: 10.1044/0161-1461(2009/09-0023)

Jones, S., Fox, C., Gillam, S., and Gillam, R. B. (2019). An exploration of automated narrative analysis via machine learning. *PLOS ONE,* 14, e0224634. doi: 10.1371/journal.pone.0224634

Kemp, K., and Klee, T. (1997). Clinical language sampling practices: results of a survey of speech-language pathologists in the United States. *Child Lang. Teach. Ther.* 13, 161–176. doi: 10.1177/026565909701300204

Klatte, I. S., van Heugten, V., Zwitserlood, R., and Gerrits, E. (2022). Language sample analysis in clinical practice: speech-language pathologists' barriers, facilitators, and needs. *Lang. Speech Hear. Serv. Sch.* 53, 1–16. doi: 10.1044/2021_LSHSS-21-00026

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting Intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

McLennan, D., Noble, S., Noble, M., Plunkett, E., Wright, G., and Gutacker, N. (2019). The English indices of deprivation 2019: Technical report. United Kingdom: Ministry of Housing, Communities and Local Government.

Miller, J. F. (1996). "Progress in assessing, describing, and defining child language disorder," in *Assessment of Communication and Language. 6th* ed (Paul H. Brookes).

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: 10.1111/j.1540-5826.2006.00205.x

Miller, J. E., Andriacchi, K., and Nockerts, A. (2019). ASSESSING LANGUAGE PRODUCTION USING SALT SOFTWARE A Clinician's Guide to Language Sample Analysis. Madison, WI: SALT Software, LLC.

Norman, K. L., and Kirakowski, J. (Eds.) (2018). *The Wiley Handbook of Human Computer Interaction* John Wiley & Sons, Ltd.

Pavelko, S. L., Owens, R. E., Ireland, M., and Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: results of a Nationwide survey. *Lang. Speech Hear. Serv. Sch.* 47, 246–258. doi: 10.1044/2016_LSHSS-15-0044

Pezold, M. J., Imgrund, C. M., and Storkel, H. L. (2020). Using computer programs for language sample analysis. *Lang. Speech Hear. Serv. Sch.* 51, 103–114. doi: 10.1044/2019_LSHSS-18-0148

Scott, A., Gillon, G., McNeill, B., and Kopach, A. (2022). The evolution of an innovative online task to monitor Children's Oral narrative development. *Front. Psychol.* 13:903124. doi: 10.3389/fpsyg.2022.903124

Stein, N., and Glenn, C. (1979). *New Directions in Discourse Processing*. Norwood, NJ: Ablex Publishing.

Tucci, A., Plante, E., Heilmann, J. J., and Miller, J. F. (2022). Dynamic Norming for Systematic Analysis of Language Transcripts. *Journal of Speech, Language, and Hearing Research* 65:320–333. doi: 10.1044/2021_JSLHR-21-00227

Wagner, R., Nettelbladt, U., and Sahle, B. (2000). Conversation versus narration in pre-school children with language impairment. *Int. J. Lang. Commun. Disord.* 35, 83–93. doi: 10.1080/136828200247269

Westerveld, M. F. (2014). Emergent literacy performance across two languages: assessing four-year-old bilingual children. *Int. J. Biling. Educ. Biling.* 17, 526–543. doi: 10.1080/13670050.2013.835302

Westerveld, M. F., and Gillon, G. T. (2010). Profiling oral narrative ability in young school-aged children. *Int. J. Speech Lang. Pathol.* 12, 178–189. doi: 10.3109/17549500903194125

Wilder, A., and Redmond, S. M. (2022). The reliability of short conversational language sample measures in children with and without developmental language disorder. *J. Speech Lang. Hear. Res.* 65, 1939–1955. doi: 10.1044/2022_JSLHR-21-00628

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

Yeung, G., and Alwan, A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. *Interspeech* 2018, 1661–1665. doi: 10.21437/Interspeech.2018-2297

Frontiers in Communication

Investigates the power of communication across culture and society

A cross-disciplinary journal that advances our understanding of the global communication revolution and its relevance across social, economic and cultural spheres.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact