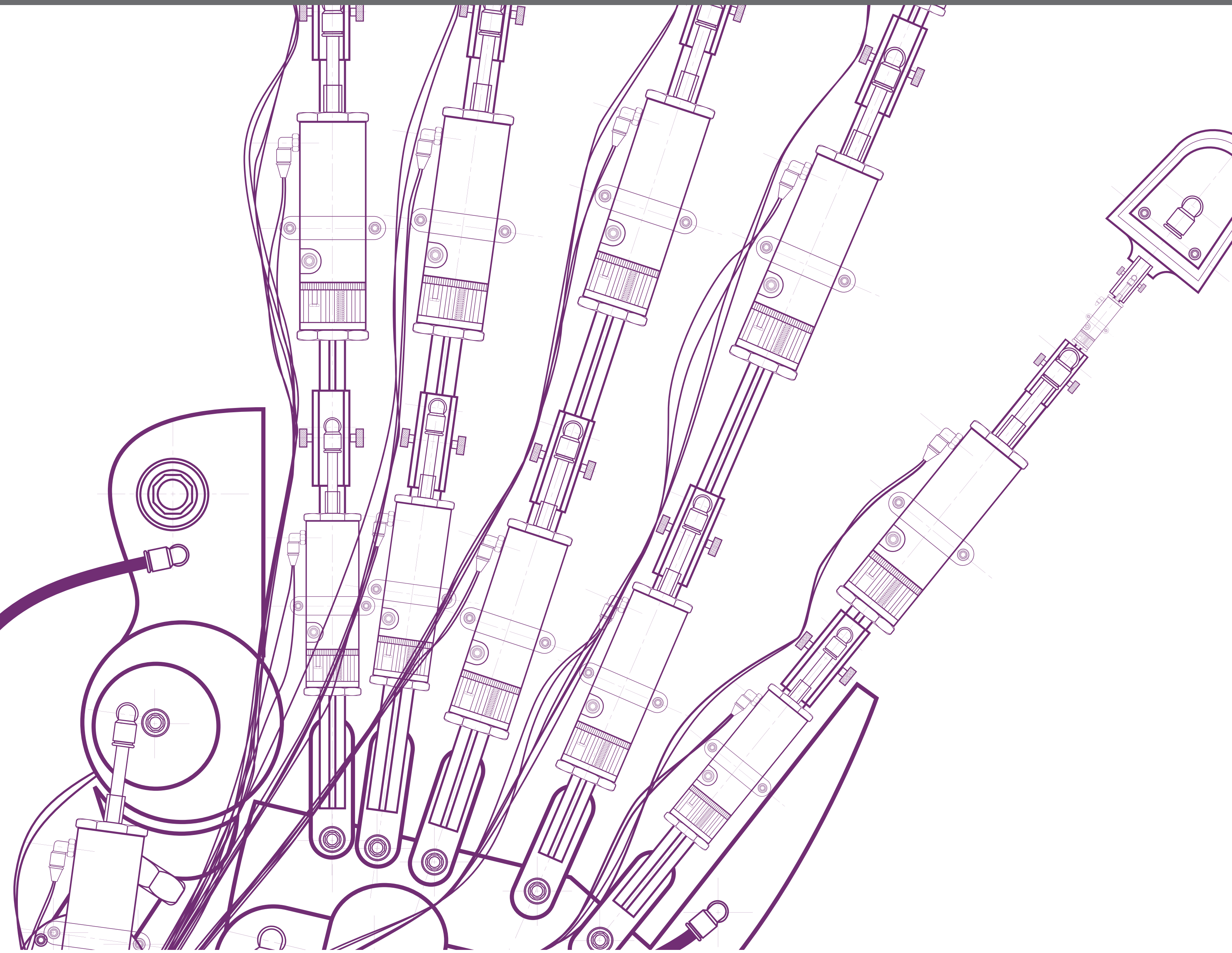
ROBOTS FOR LEARNING

EDITED BY: Wafa Johal, Tony Belpaeme and Mohamed Chetouani
PUBLISHED IN: Frontiers in Robotics and AI



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-708-7
DOI 10.3389/978-2-83250-708-7

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

ROBOTS FOR LEARNING

Topic Editors:
Wafa Johal, The University of Melbourne, Australia
Tony Belpaeme, Ghent University, Belgium
Mohamed Chetouani, Sorbonne Université, France

Citation: Johal, W., Belpaeme, T., Chetouani, M., eds. (2022). Robots for Learning. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-708-7

Table of Contents

TYPE Editorial
PUBLISHED 21 October 2022
DOI 10.3389/frobt.2022.1050658

OPEN ACCESS

EDITED AND REVIEWED BY
Alessandra Sciutti,
Italian Institute of Technology (IIT), Italy

*CORRESPONDENCE
Wafa Johal,
wafa.johal@unimelb.edu.au

SPECIALTY SECTION
This article was submitted to Human-Robot Interaction,
a section of the journal
Frontiers in Robotics and AI

RECEIVED 22 September 2022
ACCEPTED 29 September 2022
PUBLISHED 21 October 2022

CITATION
Johal W, Belpaeme T and Chetouani M (2022), Editorial: Robots for learning.
Front. Robot. AI 9:1050658.
doi: 10.3389/frobt.2022.1050658

Editorial: Robots for learning

Wafa Johal[1,2]*, Tony Belpaeme[3,4] and Mohamed Chetouani[5]

[1]School of Computer Science and IT, FEIT, University of Melbourne, Melbourne, VIC, Australia, [2]Faculty of Engineering, University of New South Wales, Sydney, NSW, Australia, [3]IDLab-imec, Ghent University, Ghent, Belgium, [4]Centre for Robotics and Neural Systems, University of Plymouth, Plymouth, United Kingdom, [5]Institute for Intelligent Systems and Robotics, CNRS UMR 7222, Sorbonne University, Paris, France

KEYWORDS

robot for education, learning with robots, human-robot interaction, personalised learning, interactive learning

Editorial on the Research Topic
Robots for learning

Introduction

In the last decade the interest in applying Human-Robot Interaction research to education and learning has boomed. While robots were traditionally used to enthuse learners about STEM subjects, it is now clear that robots can serve to teach subjects beyond STEM, such as languages or handwriting. Researchers have investigated the potential of robots to act as learning or teaching companions for children in classrooms or at home, for elderly people to help maintain their cognitive and physical abilities, and for learners with deficiencies by adapting content to their capabilities. Similarly to Intelligent Tutoring Systems (ITS), robots could have the potential to improve learning gains by personalising content and delivery to the learner. Beyond ITS, robots have a physical presence and can be used for kinaesthetic interaction and embodied learning. Finally, the social capabilities of certain robots could be used to provide adaptive empathic feedback and to engage the learner and to motivate her during the learning task.

While robots for learning is an applied topic of HRI, the context of learner-robot interaction is one of the most challenging and interesting for HRI research. Indeed the research in robots for learning often requires us to work with challenging populations (e.g., children, people with disabilities), it also requires challenging technical integration, and has very clear performance outcomes (i.e., learning gains). In some settings it even requires us to address robot-group interaction, autonomous decision making, joint attention, affective computing. Aiming to go beyond individual interfaces or projects, this Research Topic aimed to attract contributions that enable the generation of guidelines and principles for the design of learner-robot interaction.

This Research Topic focuses on social robotics research, showcasing novel algorithms and computational modeling that are applied within the context of learning. Special focus was given to contributions proposing novel theories, models, and methods for learning with robots. We also welcomed original technical contributions presenting technical systems, algorithms, and computational methods that are tailored to learner-robot interaction. In particular, we were interested in contributions demonstrating the specificities of learner-robot interaction compared to classical human-robot interaction systems.

Research Topic formation

This Research Topic emerged from a series of workshops and research projects involving the editors, Dr Wafa Johal, Professor Tony Belpaeme and Professor Mohammed Chetouani. Each in our own way, we were drawn to the challenge and promise of using robots in education, and were united in our fascination for the subject and potential to be one of the most impactful applications of social robotics.

Contents of the Research Topic

Design of robots for learning

Several stakeholders can be involved in the design of robots for learning: students, their parents, teachers or even organisations. In Lin et al., the authors look into the attitudes of parents towards storytelling robots for children. They present the challenges and opportunities of robots as storytelling companions and highlight the importance of well-being and attention for quality of life of parents when designing robot companions for their children. Johal et al. envision the role of social drones in education taking the perspective of students. They found novel opportunities for drones to potentially support group work but also several challenges that should be addressed in future research on social drone in education. In Smakman et al., the authors interviewed school teachers to examine whether social robots in primary education might compromise the social-emotional development of children, a fear often expressed by teachers and parents who have no experience with robots. Their studies indicate that robots pose little threat, and that on the contrary they might be a preferred educational technology for children with special needs.

Effectiveness and affectiveness in robots for learning

A key aspect of any educational technology is its impact on learning and learners. Not only are they designed to support teachers, but they need to show their effectiveness in allowing students to learn. In their paper, Kanero et al. use both affective and learning measures to evaluate whether a robot presented over video or a disembodied voice would be more effective at teaching vocabulary in a second language, but found no significant difference. Fischer et al. evaluated whether a robot's speaking style during second language learning mattered and found that a charismatic delivery resulted in increased learning gains.

Robots for physical learning

The fact that robots are physical devices has long been touted as important to their use in education. Most often reference is made to their tangible social presence, but robots can also offer a physical experience without necessarily offering a social experience. Kianzad et al. show how physical robots can be deployed in education and present a framework dubbed Physically Assisted Learning (PAL) for learning through haptic support.

Individual differences in learning with robots

Several studies found that there are significant individual differences in the way students interact and learn with robots. Song et al. report that the level of expertise of the learners (novice or advanced) influenced the potential role and task of a robot supporting the learning of a musical instrument. van den Berghe et al. found in their study that the background knowledge of students influenced their learning experience with a robot tutor. Tolksdorf et al. studied the influence of shyness on robot-learner interaction and found that shyness influenced the affective and the learning aspects.

Co-learning and adaptation in learning and teaching

Real-time adaptation is also crucial in educational settings. Lee et al. investigated the use of Inverse Reinforcement learning for robots to become better teachers, while Van Zoelen et al., studied how robots could adapt to become better co-leaner in a human-robot teamed search and rescue task.

Conclusion

Robotics in education, while new, offers plenty of scope and opportunity, which results in a very broad, diverse and rewarding research landscape. Testament to this are the papers collected in this Research Topic. The results presented show that there still is scope to further explore the pedagogy of using robots in education: a one-size-fits-all approach is unlikely to be effective, but how the robot should tailor its responses to maximise the learner's well-being and learning gains still remains an open question. Beyond the how and why of introducing robots in education, there remains the formidable technical challenge of building a concerted interactive experience with robots. From building robots that can robustly operate in the challenging environment of the classroom, to robots that automatically generate responses that empower the learner, the technical challenges abound. However, the promise and potential of robots in a supporting role are significant and well warrant the formidable research efforts of which this collection is but a sample.

Author contributions

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

Acknowledgments

We would like to thank Pierre Dillenbourg, who was originally part of the editorial team. We would also like to thank all the authors, reviewers and other editors who participated in this Research Topic. Wafa Johal would like to thank the Australian Research Council for their partial funding contribution (Grant No DE210100858).

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
in Robotics and AI

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/frobt.2021.579993

Parental Acceptance of Children's Storytelling Robots: A Projection of the Uncanny Valley of AI

Chaolan Lin[1]*, *Selma Šabanović*[2], *Lynn Dombrowski*[3], *Andrew D. Miller*[3], *Erin Brady*[3] and *Karl F. MacDorman*[3]

[1]Department of Cognitive Science, University of California, San Diego, CA, United States, [2]The Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, United States, [3]The School of Informatics and Computing, Indiana University, Indianapolis, IN, United States

Parent–child story time is an important ritual of contemporary parenting. Recently, robots with artificial intelligence (AI) have become common. Parental acceptance of children's storytelling robots, however, has received scant attention. To address this, we conducted a qualitative study with 18 parents using the research technique design fiction. Overall, parents held mixed, though generally positive, attitudes toward children's storytelling robots. In their estimation, these robots would outperform screen-based technologies for children's story time. However, the robots' potential to adapt and to express emotion caused some parents to feel ambivalent about the robots, which might hinder their adoption. We found three predictors of parental acceptance of these robots: context of use, perceived agency, and perceived intelligence. Parents' speculation revealed an uncanny valley of AI: a nonlinear relation between the human likeness of the artificial agent's mind and affinity for the agent. Finally, we consider the implications of children's storytelling robots, including how they could enhance equity in children's access to education, and propose directions for research on their design to benefit family well-being.

Keywords: artificial intelligence, design fiction, parent-child storytelling, social robotics, technology acceptance, uncanny valley

Edited by:
Mohamed Chetouani,
Université Pierre et Marie Curie,
France

Reviewed by:
Hagen Lehmann,
University of Macerata, Italy
Ginevra Castellano,
Uppsala University, Sweden

Correspondence:
Chaolan Lin
chlin@ucsd.edu

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: *04 July 2020*
Accepted: *09 February 2021*
Published: *19 May 2021*

Citation:
Lin C, Šabanović S, Dombrowski L, Miller AD, Brady E and MacDorman KF (2021) Parental Acceptance of Children's Storytelling Robots: A Projection of the Uncanny Valley of AI. Front. Robot. AI 8:579993. doi: 10.3389/frobt.2021.579993

INTRODUCTION

"Once upon a time" is more than an opening line to children's stories. For many, it is a tender phrase from their fondest childhood memories, suffused with parental love. Story time has life-long implications for both children and parents. Parent–child storytelling shapes the family's identity (Kellas, 2013), culture (Kellas and Trees, 2013), rituals (Fiese and Winter, 2009), and cohesion (Frude and Killick, 2011). From their first words, children learn conversational skills through turn-taking, joint attention, and the facial expressions of their parents or others conversing (Casillas et al., 2016; Casillas and Frank, 2017). In brief, children's story time is often a critical practice in parenting, enculturation, and education. To make the most of children's story time, parents have used technologies including digital books, interactive games, and talking toys.

In recent years, robots controlled by artificial intelligence (AI) have emerged as an exciting innovation for education (Fabiane, 2012; Toh et al., 2016), entertainment (Hoffman and Ju, 2014), cognitive therapy (Dautenhahn and Werry, 2004; Chang and Sung, 2013), and healthcare (Broadbent et al., 2009; Shibata and Gerontology, 2011; De Graaf et al., 2015). During the COVID-19 pandemic, the potential for deploying robots in real life was promoted (Yang et al., 2020). As Gates had predicted, robotic devices could become ubiquitous—"a robot in every home" (Gates, 2007). Thus, it is worth considering the potential acceptance and use of robots for children's story time in the home.

Domain experts have started discussions on topics including human preferences regarding decisions made by machines (Awad et al., 2018), trust in robots (Hancock et al., 2011), explanations of robot behaviors (De Graaf and Malle, 2019), and so on. However, to ensure robots fit their context of use, their design principles should be derived from the study of their intended social ecology (Šabanović, 2010). With notable exceptions (De Graaf and Ben Allouch, 2013), few studies have analyzed sociocultural influences affecting the acceptance of robotic technology for children's story time. This study begins to fill this gap.

As parental beliefs and values about children's technology inform their views on appropriate use (Eagle, 2012), our research questions were as follows: 1) To what extent do parents accept children's storytelling robots in the home? 2) How do parents envision these robots in the future? 3) What aspects of the robots could support or hinder their acceptance among parents? To address these questions, we conducted semi-structured interviews with 18 parents of children, age 2 to 5. We used design fiction, a research technique for participants to envision the use of a fictitious technology (Blythe, 2014; Lindley and Coulton, 2015).

Our main findings were as follows: 1) Despite concerns, parents were generally willing to accept children's storytelling robots. 2) Some parents viewed the robot as their replacement, a "parent double." By contrast, they viewed screen-based technologies as a way to keep their children occupied when they are busy with other things. 3) Parents valued a robot's ability to adapt and express emotion but also felt ambivalent about it, which could hinder their adoption. 4) The context of use, perceived agency, and perceived intelligence of the robot were potential predictors of parental acceptance. 5) Parents' speculation revealed an uncanny valley of AI: a nonlinear relation between the perceived human likeness of the artificial agent's mind and affinity for the agent. Two issues that could elicit cognitive dissonance were discussed: affordances of AI and mind perception of robots. Finally, we propose research directions for designing robots that enhance family wellness and meet the needs of parents and their children in everyday home settings.

BACKGROUND

Our study connects with four bodies of work. We examine children's story time with parents, review the design of story-time technology, investigate the role of parents in their children's technology use, and revisit the pros and cons of existing protocols for evaluating technology acceptance.

Children's Story Time With Parents

Storytelling is a common way for families to spend time together. It occurs as a form of family communication in either a discursive or unified fashion. Children are exposed to 1,000 to 2,000 words every hour from parents who talk as they go about their daily activities (Hart and Risley, 1999). Regular exposure to stories promotes language acquisition (Soundy, 1993), emergent literacy (Allison and Watson, 1994; Speaker, 2000), and intellectual development (Kim, 1999). Exposure to stories helps children acquire a first language while maturing and developing (Chomsky, 1972). Parental storytelling can promote reading readiness, positive attitudes, and achievement (Silvern, 1985). Through stories told orally, children acquire syntax and listening comprehension, which later support reading comprehension (Shanahan and Lonigan, 2013).

Story time helps children learn how to make sense of their experiences and relate to other people (Wells, 1986). Children's language acquisition occurs in the "social context of discourse, in the miniaturized culture that governs the communicative interaction of children and adults" (Bruner, 1981). The social nature of story time can support and extend children's social life. Stories help children develop an understanding of human behavior and the world through imagination (Benton and Fox, 1985).

According to narrative performance theory, storytelling is a way of performing family identity (Langellier and Family, 2006); family storytelling constitutes children's particular identities through content-ordering—for example, by drawing on and distinguishing social and cultural resources, such as class, race, and culture.

Bronfenbrenner's ecological model (1979) emphasizes the centrality of the family and especially the parents in a child's development. The ways children experience storytelling depend heavily on parental beliefs and involvement. For instance, parents may spend less time with their children as societal values shift toward individualism (Whitehead, 1991). Moreover, the goals parents set for their children's development can influence how they interact with them (Schneider et al., 1997).

Parent–child interactions during story time, such as turn-taking, do more to support children's language development than mere exposure to speech (Romeo et al., 2018). Additionally, parent–child attachment enhances the quality of children's involvement in story time (Bus et al., 1997). However, Bergin (2001) found that shared reading may not be beneficial if parents are hostile and critical of their children.

In sum, story time can play a crucial role in a child's upbringing. The efficacy of children's story time may differ widely because of parental attitudes and involvement, resulting in complex and unequal opportunities among children.

Technology for Children's Story Time

The use of artifacts during story time is not uncommon. In prehistoric times, parents may have told their children stories around campfires, employing props like stones, branches, bones, and so on. In ancient times, parents read stories from papyrus. In the 15th century, the printing press enabled the spread of books, which eventually led to a flowering of children's books, especially in the second half of the 19th century. In the 20th century, parents would sometimes use a record, CD player, or television program during children's story time. At the advent of the 21st century, new technologies have been reshaping children's experience. For instance, children can access storytelling with the click of a hyperlink thanks to the personal computer and the Internet.

Researchers in academia and industry have been using various technologies to facilitate and understand children's story time. For example, stories "read" by an iPod Shuffle were found to engage and motivate K–12 students (Boeglin-Quintana and Donovan, 2013). Using videoconferencing, researchers found ways to create children's story time for families separated by long distances (Ballagas et al., 2010). Interactive literature on smartphones and tablets has helped children improve their reading comprehension through role-playing (Borgstrom, 2011). Researchers have used 3D virtual narratives to explore children's understanding of stories (Porteous et al., 2017). Recently, virtual assistants like Alexa were found to engage parents with their children in story time through a voice interface (Beneteau et al., 2020).

For children's story time, one technology stands out in the progression from artifact to agent: social robots. Robotic storytellers are also part of a historical progression. The first talking dolls date back to 1890; they were made possible by the invention of the phonograph in 1871 (Plowman, 2004). In 1959, Chatty Cathy appeared as a pull-string talking doll. In 1985, Teddy Ruxpin, introduced as "the world's first animated talking toy," could move its mouth and eyes while "reading" stories played on a tape deck in its back. In 2002, Cindy Smart was marketed as the first doll that could recognize 650 words in English and some foreign words.

Despite warnings that electronic toys might inhibit children's short- and long-term development (e.g., Levin and Rosenquest, 2001), AI-enabled storytelling robots have helped children learn in various ways. For instance, robots supported children's language acquisition as learning companions in a storytelling game (Kory-Westlund and Breazeal, 2014). Affectively personalized robots can assume the role of a tutor for children's second language learning through storytelling (Gordon et al., 2016). In recent years, Codi, Trobo, and other storytelling robots have been marketed as providing developmental support outside of the classroom. Nevertheless, the influence of these robots requires investigation.

Parental Mediation of Technology Use

Novel technologies challenge most parents (Bowman, 2012). By letting children form more contacts outside the home, they can make it harder to establish family norms and sanctions (Lynd and Lynd, 1929). Although most parents consider educational robots beneficial, they lack confidence in their ability to join the child–robot interaction (Lin et al., 2012). Parents' control of technology use influences the child's development and the parent–child relationship (Giles and Price, 2008).

The way parents view technology has always been complex. For example, parents value cell phones for letting them keep in touch with their children but also worry about their effects (Boyd, 2014). Parents typically mediate their children's use of technology, including television, video games computers, and the Internet. They implement strategies like co-using and restrictions with filters and monitoring software (Livingstone and Helsper, 2008).

A parent's beliefs about children's technology use could be shaped by the parent's age, education, employment history, geographical location (Haight et al., 1999), and childhood (Plowman, 2015). Moreover, parents' and children's behavioral patterns can affect each other in various ways. Regardless of their involvement in child–technology interaction, parents provide support and guidance to their children, which in turn affect children's behavior patterns and attitudes toward technology (Lauricella et al., 2015).

Owing to the increasing complexity of the technology landscape, traditional parental mediation theories need to be revisited (Jiow et al., 2017). So far, these theories have mainly examined social and psychological media effects and information processing (Clark, 2011). A common theme in this literature is that parental mediation of children's technology use reflects their effort to mitigate its perceived adverse effects. Therefore, assessments of social acceptance of children's technology, especially social robots designed for the home, should not overlook parental attitudes and family dynamics.

Technology Acceptance

Technology acceptance denotes a user's willingness to adopt a system and that system's social and practical acceptability (Nielsen, 1993). A practically acceptable system may not be socially acceptable. Examples of social opposition include movements to ban nuclear power and genome editing. The rationale for social opposition may reflect a complex mixture of concerns, including morals, religion, political ideologies, power, economics, physical safety, and psychological well-being (Otway and Von Winterfeldt, 1982).

Researchers have been formulating various theoretical models to assess user acceptance of technology, beginning with the technology acceptance model (TAM, Davis et al., 1989). Venkatesh and colleagues compared eight prominent models to extend TAM, empirically validating the unified theory of acceptance and use of technology (UTAUT, Venkatesh et al., 2003). Due to the increasingly complicated context of use, researchers have been revising acceptance models for recent technologies, such as multi-touch displays (Peltonen et al., 2008), gestural interfaces (Montero et al., 2010), and speech interfaces (Efthymiou and Halvey, 2016). With the development of AI technologies, the ethics of adopting novel technologies are gaining more attention (e.g., Awad et al., 2018; Malle and Scheutz, 2018).

Social acceptance of robots could predict comfort with being in contact with them regularly. To succeed a social robot must be emotionally acceptable (De Graaf and Ben Allouch, 2013). Popular technology acceptance models like TAM and UTAUT were found to be limited for social robots. A study of Polish professionals' acceptance of a humanoid robot for children with atypical development found attitudes toward technology were only a weak predictor of intention to use (Kossewska and Kłosowska, 2020). De Graaf and colleagues' study showed the role normative beliefs play in the acceptance of social robots in the home (De Graaf et al., 2017).

Evaluating the social acceptance of robots is challenging. First, social robots are not merely a new form of technology. They embody human values through their humanlike presentation. In Japan, for example, robots were deployed in ways that reify "traditional" values, such as the patriarchal extended family and sociopolitical conservatism (Robertson, 2007). Moreover, various factors can challenge the validity, reliability, and practical applicability of evaluation methods (Lindblom and Andreasson, 2016). For example, the high cost of manufacturing sturdy robots that can function and survive in a home setting may compel researchers to rely on laboratory studies.

METHODS

To explore parental acceptance of children's storytelling robots, we conducted a qualitative study employing design fiction, which is a form of speculative design that opens up discussions on the use of emerging technologies and their ethical and social implications (Dunne and Raby, 2013; Hales, 2013; Malpass, 2013; Cheon and Su, 2017). This activity let parents speculate on the future of children's robots and their expectations and concerns.

Participants

Participants received an invitation by email, campus forum, local Reddit community, or word of mouth. Inclusion criteria were adult parent of at least one child, age 2 to 5. The study focuses on preschool-aged children because they are more likely than older children to be cared for at home and because robots designed for this age group are less studied. These criteria provided a new baseline for comparing robot acceptance because previous research focused on interactions between social robots and older children (e.g., Tazhigaliyeva et al., 2016).

Eighteen parents from a midwestern city in the United States and its environs participated in the study. For the method used, we believe the sample size of 18 parent–child dyads reached saturation because, in the last few interviews, no new themes were observed. Guest and colleagues found that for interview studies saturation usually occurs within the first twelve interviews, and basic elements for metathemes emerge as early as the sixth interview (Guest et al., 2006).

Fourteen participants (77%) were mothers.[1] Parents ranged in age from 24 to 38 ($M = 32$, $SD = 4$) and had a range of education levels from some college to a doctoral degree. Twelve were White, two were Black or African American, one was American Indian or Alaska Native, one was Asian, and two were another race or ethnicity. Half of the participants were full-time employees, two were part-time employees, four were students, and three were unemployed. Throughout the paper, we attribute quotes to a specific participant by using *M* for mother or *F* for father followed by a number.

This study was approved by Indiana University's Office of Research Administration (February 16, 2018, No. 1801962828). Informed consent was obtained from all participants. Study protocols complied with federal, state, and university standards, policies, and regulations.

Robots

To help parents imagine potential robot features and to inspire them to brainstorm, we searched for commercial robots that can read or tell stories. Two robots, Luka and Trobo,[2] served as probes for design fiction (Schulte et al., 2016). We selected these robots because they vary in form (zoomorphic vs. humanoid), voice (human vs. robotic), materials (hard vs. fluffy), degree of autonomy, and ability to express emotions. Luka interacts with users autonomously. It can speak several sentences (e.g., "I am bored") or blink to express emotions and attract attention. Luka has touch sensors distributed on its body. A small camera is mounted in Luka's eye area, which enables it to "read" books. Trobo, by contrast, looks like a stuffed toy though with the shape of a humanoid robot. Trobo uses Bluetooth to read e-books on its phone application. To make Trobo appear to read physical books, we used the *Wizard of Oz* technique (Green and Wei-Haas, 1985). A researcher controlled Trobo's reading pace remotely while the participants turned the pages.

Procedure

Parents provided demographic information on their family and their experience with robots via an online survey. Then, the researcher met with parents and their child at their home or in the lab.

Upon meeting, the researcher showed parents how to use the robots and asked them to use the robots for their child's story time. Parents were free to choose which robot to use first and how to be involved in the two child–robot interaction sessions. For example, some parents helped their child turn the pages of the storybook, while others encouraged the child to have story time with each robot independently. Each session lasted until the story finished (about 5 min) or was terminated by the parents when their child was too antsy or inattentive.

After the sessions, parents were interviewed for 20–50 min and were audio-recorded for transcription. During the interview, the child was given toys and a drawing kit to stay occupied. Some parents also brought a tablet computer or their partner to keep their child occupied so the interview could proceed uninterrupted.

The interviews were semi-structured. Parents were asked to reflect on their motivations, routines, and the technologies they used, if any, for children's story time. Then, we introduced the prompt for design fiction, which was a narrative of their own creation: "a robot designed to read or tell children stories for daily use in the home" (Stanley and Dillingham, 2009). Parents were asked to envision the context of use, the features of a robot they would accept, and their related thoughts (**Supplementary material S1**). During the interviews, we avoided bringing up any specific topics, such as privacy, security, and so on. In addition, we reminded parents that they were imagining a futuristic robotic concept rather than evaluating any particular robot, including the two they had just interacted with. If the parents had more than one child, they were asked to think only of

[1]In the U.S., 83% of primary caregivers are mothers (Laughlin, 2013).

[2]Manufacturer's website for Luka, https://luka.ling.ai/, and Trobo, https://mytrobo.com/.

TABLE 1 | Data themes and examples

Themes	Subthemes	Excerpt	Sample statement
Story time experiences	Approach	67	*We'll sit and read on my phone if we're out somewhere.*
	Motivation	30	*It's important for her development, for language development, imagination, and bonding.*
	Emotion	19	*I really don't like reading some of those books over and over and over again.*
	Content	17	*We tell stories from books. If we make up stories, it's just characters he knows, like daniel tiger or from PBS kids.*
Envisioned storytelling robots	Robot–child interaction	108	*It looks more like having a friend, having a second person than having just toys everywhere.*
	Robot appearance	36	*Maybe a size as big as a human being can be... if it is possible.*
	Robot intelligence	13	*If the child is talking back while they're reading, the robot should be able to interact with the child.*
	Robot-delivered content	13	*Probably, no UFO stories or anything weird.*
Attitudes toward using robots	Concern	93	*I'm afraid she would lose her interpersonal skills and knowing how to interact with humans.*
	Positive attitude	88	*If he (her child) would pay attention to the robot and sit there, then I could get something else done. That would be nice.*
	Robot-related experience	65	*. . . cause alexa doesn't understand the kids all the time.*
	Using robot vs. other technology	21	*The robot is made for a kid, and it had kid's content, whereas iPads are not just made for kids. There's lots of other stuff.*

their 2–5-year-old child throughout the study. Parents received a $25 Amazon gift card as compensation.

Data Analysis

To code the interviews, we employed grounded theory (Glaser, 1978). The first author analyzed the interview data using open, axial, and selective coding in MAXQDA (vers. 18.0.7). The other authors evaluated the trends and validity of themes. We extracted 570 excerpts from the transcripts. Through iterative memoing and refinement of categories (Corbin and Strauss, 2008), we developed three overarching themes: Story time experiences, envisioned storytelling robots, and attitudes toward using robots. Within these themes we developed further categories (**Table 1**).

FINDINGS

This section describes parents' experience of children's story time, their attitudes toward storytelling robots, their context of use, their vision for robot features, and their concerns. It also relates these descriptions to the literature.

Parent–Child Story Time

Parent–child storytelling is distinguished from other family activities by its combination of instructional value for the child's literacy and its entertainment value for both parent and child. In F2's words, story time was special because it was "*education and playing at the same time*." Parents commonly started reading for their children by their eighth or ninth month, if not earlier.

Motivation for Parent–Child Storytelling

Parents often linked storytelling to literacy education. They reasoned that story time stimulates children's curiosity, imagination, and creativity and builds their vocabulary, all of which benefit brain development and language acquisition. M9, a children's librarian, stated, "*All the brain synapses and connections are made in those first three years. So, it's really important to read to them and tell them stories and have them learn as many words as possible because their brains are like sponges; they soak up everything.*" M3, a mother of a daughter who had apraxia (i.e., motor-speech disorder), explained, "*It was hard to get her to speak because what was going on in here was processed differently. Reading out loud is very good for her because it's teaching her lots of words.*"

Some parents avoided telling their children stories because they thought it might harm their literacy. F1 explained, "*I'm really bad at phonetics, because I'm really bad about adding letters. So, I'd feel bad if he walked into school and said whatever word it was, and people made fun of him. That would stink.*" This indicates how some parents tried to balance being a help or hindrance to their child's development of literacy.

Parent–child story time was often valued as a family tradition and time for bonding. M8 stated, "*My dad did it for me, and so it kind of reminds me of that time when I was a kid and my dad was lying in bed with me reading to me. Like a cultural tradition.*" F3 explained, "*I grew up on them. It's kind of a staple of growing up, having stories read to you by your parents.*" Story time builds closeness and is one of the joys of parenthood. M11 said, "*We cuddle, so we sit close together, and I'll put my arm around her, and it's just kind of a bonding time.*" Some parents simply enjoyed story time. F1 stated, "*It's fun. We kind of lie in bed together and look at the pages and talk about it.*" In other words, watching children get excited about stories and learn things was a treasured part of parenthood:

> *I like watching the look on her face, 'cause sometimes she's confused, and I see her eyebrows go up and down and see her cock her head to the side. I just enjoy seeing her reaction and how curious she is. She's curious, cocks her head, moves her eyebrows. It's like I can see the wheels turnin'. I don't know what they're doing up there, but she's definitely thinking.* [M14]

Some parents use story time to set up their child's evening routine. As part of a bedtime ritual, parents use storytelling to help the child regulate affect (e.g., to settle down). M3 stated, "*When it's story time, she's not up running around and playing. And she knows that story time is the progression in going to bed.*" This ritual could also give children a sense of being part of a family and help to maintain the parent–child bond (Franklin and Bankston, 1999). Other parents hoped that their children would form healthy long-term habits through an early attachment to books. M14 explained, "*We want her to be curious and learn, and we know that it starts at a young age. So, by reading stories to her now, we hope she's going to continue to want to read and continue to want to learn.*" These examples reveal storytelling as a way of parenting.

Stories took various forms. Some parents read fairytales or short stories because they fit their children's attention spans. Other parents told family stories. Their children enjoyed hearing about what happened before they were born. In particular, parents tended to involve younger children in the storytelling process, such as setting a scene, developing characters, creating a plot, and so on. In other words, they created stories with their children, instead of for them. This made story time a venue for self-expression as well as social interaction, which situated the children as active agents in constructing their sense of self (Korn, 1998):

> *Sometimes, she'll say, 'Mom, tell me a story,' and I'll say, 'Okay, once upon a time there was a princess named Charlie [her daughter's name],' and then I'll say, 'And then what happened?' and then she'll say, 'Oh a big dragon came and took her away,' and so we just kind of create stories doing that. As she got a little older, we would read books more* [M11].

In sum, parent–child storytelling could serve multiple goals, involve dynamic interactions, and nurture mutual well-being. Balancing its contribution to a child's literacy with potential adverse effects could pose a challenge to some parents.

The Role of Technology in Story Time

Most parents preferred the use of physical books to technology. M4 stated that, although using physical books is "*old-fashioned,*" it is something children "*have to get used to, though things are becoming more and more digitized.*" A recurring reason screen-based devices were not favored was that parents doubted young children could benefit from stories played on such a passive medium. M14 explained that "*at her age, she's having fun [playing with a smart device], but she's not learning anything.*" She emphasized that during story time children should physically "*experience things*" like turning pages. Similarly, F1 observed that his son didn't look at the pictures in the book while listening to the story on a phone application. In this case, F1 mentioned, "*I try to make him pause and look at the scene that's on the page and kind of get an idea of reading comprehension, I guess, so that these words go with this picture.*" Indeed, brain connectivity in children increased with time spend reading books and decreased with time spend using screen-based media (Horowitz-Kraus and Hutton, 2018).

Nevertheless, some parents valued smartphones and tablet PCs such as Kindles and iPads for providing quick access to a large supply of reading material. M8 stated, "*We do actually read on my phone a lot. We've got one of those Kids Zone type apps, and it's got different books in there. We'll sit and read on my phone if we're out somewhere. It's just easier than dragging three or four books around.*" Some parents used multimedia, such as online videos and TV programs, as a replacement for story time to relieve the stress of parenting. M9 explained, "*The screens are the only thing that can take his attention to the point where he won't keep asking me questions. Because he's an only child, it's just me and him in the house. And he wants to interact with someone.*" M5, an exhausted mother, said, "*I don't have time to tell him a story. Those programs are already on the TV. Basically, I would rather have him once in a while sit down and listen to stories.*" These accounts reflect how parents employed technology in "digital parenting" (Mascheroni et al., 2018).

Positive Attitudes Toward Children's Storytelling Robots

This section reports parents' vision of children's storytelling robots and the robot's context of use. Parents generally held positive attitudes toward the robots, especially compared with screen-based technologies.

Parents' positive attitudes were exhibited by their vision of the robot's role and suitability for children. Parents envisioned a storytelling robot as a "parent double." For instance, M4 stated, "*If I had an especially busy night, and my husband wasn't home, and it was time to do story time. You know, that would maybe give me 10 min to pack their lunches while they listen to their story from the robot.*" Moreover, parents reported a desire to have a storytelling robot deal with "*boring tasks*" like rereading a book. M12 stated, "*If she gets to a point where she does want to do the same book, you know. Finish it and start right back over, a robot would definitely do that. I would not want to.*" Indeed, developmental studies suggested that requesting repetition in book reading is common for young children (Sulzby, 1985). Having the same book read to children repeatedly can increase their enjoyment and helps them learn new words (Horst, 2013). M10 also expected a robot to do something she didn't enjoy doing during story time: "*My older one is almost doing chapter books. It would be awesome if [the robot] read chapter books because that's what I hate—reading out loud.*" These examples show how parents value a robot contributing to children's story time and reducing their stress.

Some parents wished to delegate tasks involving emotional support. M4 explained that when children need to study, "*it would be nice to have an automated thing that could do that with them, and say, 'Hey, that's right!' or 'No, that's wrong.' You do the boring stuff, robot, and I do the fun stuff.*" Surprisingly, some parents wanted the robots to cover difficult topics like sex and death. M3 remarked, "*She was just watching Daniel Tiger, and they were talking about death, so [the robot] could cover topics that are hard to discuss [or] at least start the conversation.*" These findings show the need for robots to perform social support.

Parents thought the robot could facilitate children's story time in many ways, most of which relate to their perception of a robot as social. M13 explained, "*[The robot] would probably keep [my son's] interest a bit longer. He might think, 'It's a person. I have to stay here because it's reading to me.' He might be more fascinated with it too.*" Parents speculated that robots would engage children in social interaction while reading: "*an advantage is that your child is hearing somebody else talk.*" Parents envisioned robots acting as an educational peer. M3 observed, "*I could see that being useful when they're learning to read just because it would be a buddy to practice with, and it could maybe help her if she got stuck on a word.*" Similarly, M4 looked forward to robots that could reinforce her children's foreign language learning at home. F1, who was not confident with phonetics, said, "*If I'm stuck on a word going, 'I'm not sure buddy,' then we could be with the robot and put him in front of it and he could read the paragraph. That would be really nice.*" Some parents suggested that robots could be an authoritative social mediator:

> *Sometimes, when I'm reading with my five-year-old, she doesn't believe that a word is pronounced the way that I'm pronouncing it, and so I have gone on dictionary.com and had to play it for her. And, I'm like, 'See? Right here.' And, she's like, 'No, no, no. You're just doing that.' So, I'm kind of like, 'Well, see the robot said this is how you pronounce the word.' And, she might be like, 'Oh,* okay.' [M10]

Parents indicated that robots would be more child-friendly than televisions, tablets, smartphones, and other devices because a robot's predetermined content was controllable and trustworthy. M4 explained, "*I would kind of trust a pre-programmed thing made for kids, whereas something like YouTube, they could go down a wrong path and see things that they shouldn't be seeing.*" As such, parents preferred a robot that, unlike today's Internet (Nikken and Jansz, 2014), would not require their direct supervision. Some parents envisioned interacting with a robot would benefit children's development by reducing their screen time, which is "*bad for their eyes*" (M10). Moreover, robots were seen as addressing usability issues that impede children's technology adoption. M13 mentioned, "*He gets upset because he doesn't understand how to control the phone. A robot would be easier for kids to interact with.*" This example shows how parents envisioned it being immediately apparent how to use the robots. In other words, they expected them to have highly salient affordances.

Children's Storytelling Robots: Expectations and Concerns

Despite the perceived usefulness of children's storytelling robots, we observed a series of technical and social challenges that would affect parental acceptance and adoption. Previous studies indicate that people expect robots to have social traits that help them empathize with people (Breazeal, 2004). The present study found that parents expect storytelling robots to have social intelligence. However, a mechanical robot possessing this human quality might give parents cognitive dissonance. This psychological pain arises from inconsistent cognitions (Festinger, 1957), such as perceiving the robot as a social being while knowing it is just a machine. We identified two key factors impacting cognitive dissonance: a robot's 1) adaptive capability and its 2) affective capability. Parents expected a storytelling robot to be competent at both, though their acceptance of storytelling robots could be hindered by ethical concerns and by the uncanniness of robots that seem to be electromechanical yet possess conscious experience.

Adaptive Capability

We define *adaptive capability* as the ability to adapt autonomously to a real-world context. Parents doubted whether the adaptive capability of robots could meet the challenges of children's storytelling. A major concern was impromptu conversations between the child and the robot. In particular, parents expected robots to respond to questions from children automatically. Enhancing the robot's autonomy would likely increase its perceived usefulness and thus acceptance (Thrun, 2004).

Prior research indicates that conversational interactions during story time lead to children's literacy success (Berk, 2009). These interactions happen naturally during parent–child storytelling. For example, as M6 explained, "*When you are reading to children, they want to talk. They will not just sit and not talk as you're reading. Most times, they want to talk like 'Oh am I flying? Am I ...?'*" Thus, she envisioned the robot interacting in real time to help children engage with the story. "[*Children*] *usually have questions when they are reading, and if the robot is not answering the questions, they can't even think about their questions and all.*" Discussions could help children interpret the story, including the character's facial expressions. M13 suggested, "*The discussion is definitely important because she needs to be able to look at somebody and know if they're angry or if they're sad.*" Parents talking to their children would help them develop literacy and a love of reading (Burns et al., 1999).

Additionally, parents mentioned that a robot would need to recognize a toddler's voice. M4 said, "*We have an Echo up there, and so, a lot of times when we ask Alexa questions, she's like, 'Oh, I don't understand you. I don't understand you,' especially with the kids.*" She further explained that a child's voice is "*so different and high-pitched, or they don't pronounce words right, so I could see that being really frustrating for a kid if the robot's not understanding them.*" The robot would need a system for understanding a child's speech.

Another challenge was whether storytelling robots could maintain children's engagement to guide their attention. In real life, parents often use linguistic skills, emotional expression, and gestures to increase engagement. For example, M5 added rhyme so that the story is "*not going to be so boring*" and to keep her child on track with the story. She insisted that the skill of storytelling is unique to "*human beings.*" Some parents upheld that humans were naturally better storytellers than robots because they can gauge the child's reactions and decide what information is important to convey. Another common concern was that the robot's synthesized voice with its flat intonation and monotonous rhythm would hamper reading comprehension. Indeed, a storytelling robot's intonation and emotion predict concentration and engagement (Kory-Westlund et al., 2017).

Paradoxically, some parents preferred a robot with a low level of autonomy to converse with their child on security and ethical grounds. Parents feared that children trusting the robot could put them at risk if someone hacked the robot or recorded their conversations. F3 noted, "*If the robot was like, 'What's your dad's social security number?' I might freak out*." F2 pointed out that some questions (e.g., sexual orientation) were too sensitive for young children to discuss, even with parents. M7 thought a robot with a high level of autonomy would be threatening: "*I don't want something that's going to take over my house. I don't want her to become reliant on a robot.*" Given these concerns, parents suggested that stories requiring limited turn-taking would befit robot–child storytelling. Specifically, alphabet books and nursery rhymes are good candidates because they are easy to follow without a back-and-forth conversation.

A few parents found the idea of a robot talking with a human spooky. M9 remarked, "*It'd be like 'Hello Emily, how are you?' And I'd be like 'No! Stop talking to me.' But it's just because I watched too many scary movies when I was a kid, where things that weren't supposed to talk started talking*." Some parents recounted their impressions of early talking toys:

> *You remember those Furbies? Those things would just start talking out of nowhere, and it scared people. If the robot could just wake up and start telling a story in the middle of the night, if it started talking on its own out of nowhere, I think I would be scared of it.* [M8]

Some parents wondered how storytelling robots could flexibly adapt to complex surroundings. M14 noted that a typical story time for children would be at bedtime in dim light. The robot may be unable to "see" the book. Or, if a robot could only read in the sunroom, it could cause frustration because the sunroom might be occupied. The children might have to choose between being with the robot or being with their parents, and "*sometimes they just want to be close to mommy and daddy. They don't want to be left alone*" [M14]. In other words, the adaptive capability of a robot could affect its actual use and even make humans accommodate to it.

Affective Capability

We define *affective capability* as the robot's ability to express emotion appropriately, to influence the user's emotional engagement with the robot and topic or story, and to arouse the user's affection. To these ends, parents expected the robot to employ various features, such as appearance, emotion contagion, empathy, and social engagement.

Parents generally expected the appearance of a storytelling robot to be anthropomorphic, for example, "*having head and body, such as R2D2 in Star Wars*" [F3]. Some parents thought having arms and legs could differentiate a robot from other devices. In particular, robots with eyes were believed to help children take their reading time seriously. M5 reasoned, "*I like the fact that the robot has got eyes, because it looks like, 'Okay, we are looking at each other, so what are we talking about?' And then you could say, 'No, that's not a toy.'*" M11 expected the robot to have hands to hold up a physical book and turn pages so children "*have to stay engaged*." Parents tended to imagine the robot resembling a human to support social engagement.

However, some parents felt threatened by the idea of the robot having a humanlike appearance because the child might treat it as an alternative parent. M11 was concerned that a child would think a humanoid robot was the one "*to get reading from or to spend time with.*" M13, who thought of reading as mothering her child, explained that if the robot were not humanlike, she wouldn't feel it was taking her spot. Furthermore, several parents pointed out that robots would be creepy if they were too lifelike. One mentioned Teddy Ruxpin, an animatronic toy in the form of a talking bear: "*It's just trying too hard to be human. I would like the toy to acknowledge that it's a toy. I would want it to have some kind of a toy appearance*" [M12]. In particular, a parent expressed her eerie feeling when Teddy Ruxpin rolls its eyes and moves its mouth while talking:

> *They were scary. I never wanted to own one, because it was just like these big eyes that moved around and this tiny mouth. It was like 'Hi, I'm Teddy Ruxpin.' And I was like, no you're possessed by the Devil.* [M9]

Parents expected storytelling robots to act expressively. They emphasized that the manner of expression was critical. In M5's words, "*Robots should be able to express stories while they are saying them.*" This aligns with Bauman's "performance-centered conception of verbal art," which holds that "the formal manipulation of linguistic features is secondary" (Bauman, 1977). Specifically, parents envisioned storytelling robots as being able to express emotions through body language, gestures, eye gaze, and so on. Such communication would help make concepts easily understood by young children. Children should not be "*just like sitting, watching something on a screen, or just like listening to a tape play*" [M9]. A typical example would be indicating with the hands an object's size or shape.

Moreover, parents anticipated that emotions serve shared psychological and physiological functions, which are critical in the context of storytelling. For example, humor can hook children into story time, enhance learning, spark social interactions, and establish rapport (Savage et al., 2017). However, some parents indicated that, because humans and robots lack a shared social grounding, they would not be able to share in each other's emotions. F3 reasoned that humor could give different people different perceptions. He provided an example where a robot was reading a picture book where characters were waving long arms: "*He can't necessarily ask the robot why the arms are so long and squiggly or laugh with it that the arms look squiggly and funny. That's something that humans would be like, from one person to another, subjective, but long, squiggly arms are funny. So, you can't really have that emotional interaction with the robot.*" In sum, an emotional exchange requires social grounding, which remains a nearly insurmountable challenge for social robotics.

The affective capability of the robot led to another struggle concerning the relationship between robot and child. Some parents wanted the robot to have the affective capability to engage children in social interaction beyond story time. Unlike another device that might just be left on a shelf, they proposed children could carry the robot around, talk with it, and cuddle with it while going to sleep. However, parents worried whether interacting with the robot would impede the development of social skills or cause social dysfunction. M7 was reluctant to use the robot "*because I'm afraid she would lose interpersonal skills and knowing how to interact with humans.*" M11 highlighted human–human interactions: "*that cuddling, that hug, I think that's important when they're young.*" Parents commonly noted that, for young children to learn social cues and how to engage with others, real people were irreplaceable.

The idea of social robots that simulated human warmth perturbed some parents. In particular, some busy parents worried that leaving children to such robots might create a gulf between parent and child. The child might prefer the robot to the parents and spend more time hugging and cuddling with the robot. In this case, young children could switch their attachment from their parents to their robot. M4 worried that if she left her son with a robot, "*he might just get addicted to it and want to spend more time. He might find the robot's stories more interesting than your stories, even when you have the time to tell him stories.*" M5 worried that using a robot could reduce time spent with children and chances to get to know them: "*So, they may have imaginations that you cannot know because you are not the one telling them the story.*" These parents viewed affectively capable robots as a threat to parenthood.

Some parents disliked the idea of children's storytelling robots. They considered story time a unique part of parenting that should never be handed over to an AI-enabled robot. F1 explained, "*I think [story time] is my bonding time, this is my time to spend with my kid. So, I don't want people to use them to separate themselves, because I feel that raising a kid is very personal.*" M9 insisted that she would never give up any story time with her child, "*I love telling stories to him, and I love reading books to him. It is my job at home.*" F4 discussed stories as a way to shape his children's morality, stating, "*I don't want to get to the point where robots are informing my child on topics. It's a parents' job to parent and be responsible for their kid. I don't ever want that responsibility to be on a school or a robot or anything but my wife and me.*" These quotes underscore how some parents consider parent–child story time to be an activity exclusively for family members.

DISCUSSION

In this section, we discuss parental acceptance of children's storytelling robots in the home. We argue that parents' expectations and concerns reflect an uncanny valley of AI, which may be interpreted using the metaphor of the *castle in the air*. Finally, we explore the implications of designing children's storytelling robots and propose directions for future research.

Parental Acceptance of Children's Storytelling Robots

Our findings indicate that, despite reservations, parents would generally accept storytelling robots in their home. Their acceptance relates to how they valued children's story time. Parents emphasized their three main goals: literacy education, family bonding, and habit cultivation. Correspondingly, parents valued storytelling robots for pedagogy, felt they could threaten parenthood, and struggled with their potential effect on child development through daily use.

Parents also viewed parent–child story time as personally fulfilling and beneficial to their family. This explains the reluctance of some parents (e.g., M9, F1, and F4) to use a storytelling robot: It might steal from family time and weaken family cohesion. This concern is not without merit. Previous studies indicate task persistence reinforces attachment (Bergin, 2001). In other words, family cohesion increases with the time parents spend telling stories to their children. Family bonding during parent–child story time includes talking, cuddling, and joint attention, all occurring in a physically and socially shared space. Beyond giving comfort, parent–child touch could enhance prosocial behavior. To replicate this between child and robot is challenging (Willemse et al., 2017). In addition, storytelling gives parents a chance to start discussions that teach their children values. Thus, it supports family cohesion and shapes family identity.

According to parents' narratives, a storytelling robot may require intentional agency, which assumes a strong AI position (Searle, 1980). The robot might then be able to establish an intricate microsocial environment for human children in the home. Although some parents viewed this conception as utopian (Segal, 1986), the context of children's storytelling touched a nerve. The theory of *Ba* proposes that a living system maintains self-consistency by the contingent convergence of the separated self and the non-separated self (Robertson, 2007). Here, futuristic child–robot storytelling is a *Ba* that involves a dynamic tension between a roughly human storyteller and a developing human child. Some parents seemed disturbed by the thought of a robot guiding a child through emergent, uncertain states of development. Their concern runs counter to the expectation that a storytelling robot serve as parent double. These conflicting cognitions elicit the psychological discomfort of cognitive dissonance. Storytelling as the context of use could be a critical factor in parents' ambivalence regarding robots for children.

We interpreted a parent's expectation as reflecting different perspectives on a robot's ontological status—whether it could exhibit human likeness, agency, and emotions. In the storytelling context, parents spontaneously imagined the robot as anthropomorphic and anthropopathic. Parents tended to imagine child–robot interactions as mirroring human–human interactions such as turning the pages of a physical book. By contrast, screen-based technology typically involves a graphical or voice interface. These interfaces are often not child-friendly, which frustrates parents (e.g., McFarlin et al., 2007). Thus, a robot was deemed more suitable. Parents went on to envision scenarios where the robot acts as a peer or mediator. Parents thought a physical robot could engage socially with children and forge a relationship with them, which could benefit children's overall development. Thus, parents' perception of robot agency heightened their expectations of the future of child–robot interaction relative to other technologies.

However, some parents were also disturbed by the scenario of a robot simulating human interactions. Nevertheless, parents seemed optimistic about the ability of futuristic robots to reduce their parenting stress by serving as a "parent double" in performing "boring" or "difficult" tasks. They expected a robotic storyteller to simulate a human storyteller's physical autonomy and social intelligence. Successful storytelling would involve verbal and nonverbal interactions laden with affect. In other words, to be effective, a storytelling robot needs to respond dynamically to young children, whose communication involves various resources, such as gestures, vocalizations, facial expressions, body movements, and so on (Flewitt, 2006). However, parents' ambivalent, paradoxical feelings may have been sharpened by the robot's perceived intelligence. They struggled with its adaptive and affective capability. Parents worried children would trust the robot and follow its instruction, which could be a security threat if, for example, another person took control of it. Thus, perceived intelligence could be a third influential predictor of parents' acceptance of storytelling robots.

In sum, the context of use, perceived agency, and perceived intelligence of a robot were promising predictors of parental acceptance. Designers of children's storytelling robots should consider these factors in the design and evaluation process.

A Projection of the Uncanny Valley of Artificial Intelligence

We argue that the two factors impacting parents' cognitive dissonance, a robot's adaptive capability and its affective capability, are a projection of the uncanny valley of AI. Why did some parents envision a children's robot telling stories in a humanlike way—flipping storybook pages, pointing out illustrations, acting out scenes, and responding to disinterest or spontaneous questions—yet preferred the robot to have a low level of autonomy? Why did some parents feel weird when greeted by a robot but not when it told their children stories?

Consider the Chinese idiom *castle in the air*. It denotes the impractical dream of building a magnificent third floor before the first two floors are complete. Imagine building a three-story castle of a robot's intelligence. The ground floor is weak AI. The robot combines bottom-up processes, each designed for a particular task. For example, a robot obeying the command *read a story out loud* or *open the window* might just be simulating some disconnected aspects of human behavior. The next floor is strong AI. The robot has—or at least simulates—general human intelligence. It can apply top-down processing to figure out what to do in new situations. For example, the storytelling robot may be able to infer disengagement when the child responds slowly in a low voice. The top floor is social intelligence. The robot creates the feeling of being in the presence of a living soul—with free will or whatever being human entails. For example, if the child lost interest in a story, the robot would be able to respond like a real person. Science fiction has dramatized the top floor. In the 2001 film *AI. Artificial Intelligence*, David, the robotic boy, felt desperate about his human mother abandoning him and set out to find out why.

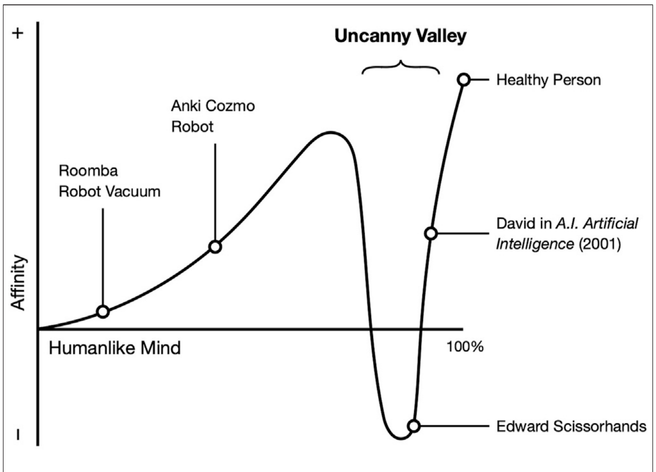


FIGURE 1 | The Uncanny Valley of AI.

As the floors of the castle of a robot's intelligence are constructed, and as its human characteristics increase, human perception begins to apply a model of a human other to the robot. If the bottom or middle floor is perceived as incomplete, the top floor becomes a castle in the air. It is unconvincing and even creepy. These perceptions reflect an uncanny valley of AI: a nonlinear relation between affinity and the perceived humanness of an artificial agent's mind (**Figure 1**).

Our proposal is a bit different from Mori's original concept of the uncanny valley of human likeness—that is, outward resemblance (Mori, 2012). Specifically, as the intelligence of an artificial agent increases, it becomes more likable, up to the point at which its perceived intelligence begins to approach human intelligence. For example, an embodied storytelling robot capable of conversation while flipping pages, is more appealing than Roomba, the cleaning robot, because the former is capable of social interaction. However, when a robot's intelligence seems more human, but is still distinguishable from human, it could elicit an eerie feeling. This could explain why some parents preferred a robot with a low level of autonomy. For a more extreme example, Edward Scissorhands' human intelligence is betrayed by his atypical way of thinking and behaving, which makes audiences see him as uncomfortably deviant (Clarke, 2008). This uneasy feeling disappears when a robot's perceived intelligence becomes indistinguishable from a real human (e.g., when an AI passes the total Turning test, Harnad, 1991; Saygin et al., 2000; Turing, 1950).

Researchers have proposed different explanations of the uncanny valley (for a review, see Kätsyri et al., 2015; Wang et al., 2015). Most theories focus on how imperfect human appearance or movement triggers eeriness (Ishiguro and Dalla Libera, 2018; Paetzel et al., 2020). For example, when a robot's human resemblance exceeds a certain point, the expectation of human performance eclipses the robot's ability to perform (MacDorman and Ishiguro, 2006). Indeed, the idea of androids alarmed some parents (e.g., M9 and M12).

However, just as a real human being can be evaluated from the standpoint of mind or awareness, a robot's intelligence could play a role distinct from its appearance. People might be unsettled by AI-enabled voice assistants like Amazon's Alexa, although the shape of the device is just a black cylinder (Thakur, 2016). Factors other than a humanlike appearance influence mind perception (Gray et al., 2007; Gray et al., 2012). To extend Mori's observation, the uncanny valley of AI predicts that 1) a certain level of intelligence facilitates social interaction between humans and robots and 2) artificial intelligence that is similar to, but still distinguishable from, human intelligence could create the uncanniness of a castle in the air.

Perceptual issues with the first two floors of a robot's intelligence could lead to this. For example, why did a parent (M8) mention that it was creepy for Furbies to suddenly start talking? There could be at least two reasons. One is the lack of transparency of the first two floors (Kory-Westlund et al., 2016; Wallkötter et al., 2020). The affordances for engaging with the robot's intelligence are unclear. For example, ordinary people have a limited understanding of the structure of the second floor (i.e., the robot's perceived capability for top-down processing), which makes it difficult for them to establish a mental model of how a robot with a social capability operates (i.e., the top floor). As such, robots with the appearance of social intelligence could create an illusion. When our brain tries to falsify the illusion but fails, our expectations falter, our brain's prediction errors accumulate, and our feeling of a social connection with that robot oscillates between what Quinton (1955) called perceptual presence and pure thought. One practical way to relieve a user's weird feelings about a robot with high intelligence should be to make its AI understandable (Wang et al., 2019).

The other issue underlying the creepy feeling of a robot that suddenly starts talking could relate to mind perception, namely, the eerie feeling caused by the attribution of mind to a machine (Gray and Wegner, 2012; Appel et al., 2020). People may perceive mind along two dimensions: *experience*, the capacity to feel and sense, and *agency*, the capacity to act and do (Gray et al., 2007). In our findings, parents' linguistic use reflects these two dimensions of mind: adaptive capability relates to agency and affective capability relates to experience. A robot with weak AI is perceived as being low in experience but high in agency. The increasing perception of a robot's experience and agency tend to reinforce each other, creating a halo effect (Nisbett and Wilson, 1977).

For example, when a storytelling robot starts talking spontaneously, perhaps merely due to a bug in its program, it can create an illusory experience: The robot appears to be more than it is (i.e., the ground floor where a robot tells stories). Contradictory perceptions and cognitions cause cognitive dissonance. The robot seems to have the capacity to act and to do something unknown (i.e., the top floor). Parents are especially unsettled by unpredictable actions as they relate to their children. Because the middle floor of the robot's intelligence (i.e., strong AI) does not yet exist, people construct the top floor as a castle in the air, which is unnerving. However, more empirical research is needed to examine the interaction between a robot's perceived experience and agency.

Children's Storytelling Robots: Implications and Future Directions

Although the development of robots as a parent double still faces technical and design challenges, their social and economic value is clear. Not every child has a caregiver with leisure time for storytelling. Across the globe, we find social crises involving children: children in orphanages and other institutions—and sometimes refugee camps—without parental love and nurturance; children in foster care, perhaps separated from abusive or neglectful parents; children whose parents are illiterate, blind, deaf, or mute; children with autism who find it easier to interact with a robot, and so on (Scassellati et al., 2012). While artificial love may never replace human love, storytelling robots could lessen inequality by simulating parental warmth during early development.

The prospect of leading educational activities in the home causes some parents stress (Deniz Can and Ginsburg-Block, 2016). Alternatives, such as having relatives or babysitters read to their children or placing their children in literacy programs, may raise issues of trust or pose a financial burden. Sometimes a human assistant may be unavailable, such as during the lockdown period of a pandemic. Thus, robots acting as a parental double during story time could help relieve parental stress.

Using storytelling robots to address the social crises mentioned above is not whimsical. Robots have been used to address social crises elsewhere. For example, Japan's government identified robotics as a solution to its looming demographic crisis caused by a lack of young people to care for older adults. Robots are also used in Japan to care for children, to provide companionship, and to perform chores (Robertson, 2007). Robots were preferred as home healthcare workers to Asian foreigners. They were considered less likely to violate cultural norms or interpret history in a way that could cause conflict (Robertson, 2007). However, social acceptance of robots could vary with religious and cultural history, personal and human identity, economic structure, professional specialization, and government policy (MacDorman et al., 2009). Thus, crosscultural issues should inform future studies on the acceptance of children's storytelling robots.

Nevertheless, envisioning a storytelling robot has raised concerns (e.g., for M9, F1, F4). For example, a robot could become a threat to parenthood or parental identity if a child shifted attachment to the robot. A young child could develop a closer relationship with the robot through even boring activities like reading a book repeatedly (Sharkey and Sharkey, 2010; Kory-Westlund et al., 2018). The embodiment and voice of a robot could be powerful indicators of social presence (Reeves and Nass, 1996). Therefore, interacting with a robot during story time could give a young child the illusion of rapport (Turkle, 2007). However, child–robot rapport is unlikely to threaten child–parent attachment. Children are predisposed to be attached to their mother, attachment has survival value (Bowlby, 1977), and begins *in utero* (Sullivan et al., 2011).

Future directions for designing children's storytelling robots include research on how to create educational and affective experiences for at-risk young children, how to promote the well-being and quality of life of parents, and design principles for healthy human–robot relationships (MacDorman and Cowley, 2006; Miklósi et al., 2017). Moreover, as both children and parents are stakeholders, their individual differences and interaction patterns could predict the success of storytelling robots. One critical variable is parenting style, which correlates with children's technology use (Chou and Fen, 2014). In addition, to evaluate parental acceptance of children's storytelling robots more accurately and to explore how the robots would be brought into the family, longitudinal studies are needed. Finally, more generalizable studies to support the proposal of the uncanny valley of AI could come from future work, including surveys, replications with a broader sample, and laboratory experiments.

CONCLUSION

The present exploratory study investigated parental acceptance of storytelling robots for young children in the home, a subject that has received scant attention. Using design fiction as a research technique, we found that household storytelling robots are more than a new type of technology for children. They provide an intricate testing ground for studies on cognitive perception, family dynamics, and human–robot interaction design.

Our findings showed that parents had ambivalent though generally positive attitudes toward storytelling robots and were willing to accept them in the home. Parents valued storytelling for their child's literacy education, habit cultivation, and family bonding. These goals provide a framework for assessing the usefulness of storytelling robots. Likely predictors of robot acceptance include context of use, perceived agency, and perceived intelligence. Parents both valued and felt concern about the robot's adaptive and affective capability.

We discussed possible mental models and cognitive mechanisms behind parental expectations. Unlike screen-based technologies, parents could see a storytelling robot as a parent double, which could relieve them of boring and stressful aspects of parenting but could also threaten parenthood. We also introduced the concept of an uncanny valley of AI to explain some of the parents' ambivalent views. Parents found it difficult to establish a mental model of how a robot with a social capability operates, which creates cognitive dissonance and a feeling of uncanniness. This feeling might be mitigated by making its AI more transparent and understandable. Finally, we explored the implications of using robots for children's story time, including their potential influence on parental well-being, and suggested directions for future research.

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 Indiana University Office of Research Administration. Written informed consent to participate in this study was provided by the participants' legal guardian or next of kin.

AUTHOR CONTRIBUTIONS

CL: Conceptualization, Methodology, Validation, Resources, Investigation, Data Curation, Formal analysis, Writing – Original Draft; SS: Writing – Reviewing and Editing; LD: Writing – Reviewing and Editing; AM: Writing – Reviewing; EB: Resources, Supervision; KM: Validation, Writing – Reviewing and Editing, Supervision.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.579993/full#supplementary-material

REFERENCES

Allison, D. T., and Watson, J. A. (1994). The Significance of Adult Storybook Reading Styles on the Development of Young Children's Emergent Reading. *Reading Res. Instruction*. 34 (1), 57–72. doi:10.1080/19388079409558171

Appel, M., Izydorczyk, D., Weber, S., Mara, M., and Lischetzke, T. (2020). The Uncanny of Mind in a Machine: Humanoid Robots as Tools, Agents, and Experiencers. *Comput. Hum. Behav.* 102, 274–286. doi:10.1016/j.chb.2019.07.031

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., et al. (2018). The Moral Machine Experiment. *Nature*. 563 (7729), 59–64. doi:10.1038/s41586-018-0637-6

Ballagas, R., Raffle, H., Go, J., Revelle, G., Kaye, J., Ames, M., et al. (2010). Story Time for the 21st Century. *IEEE Pervasive Comput.* 9 (3), 28–36. doi:10.1109/mprv.2010.45

Bauman, R. (1977). *Verbal Art as Performance*. Prospect Heights, IL: Waveland Press.

Beneteau, E., Boone, A., Wu, Y., Kientz, J. A., Yip, J., and Hiniker, A. (2020). Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics. in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–13.

Benton, M., and Fox, G. (1985). *Teaching Literature: Nine to Fourteen*. Oxford: Oxford University Press.

Bergin, C. (2001). The Parent-Child Relationship during Beginning Reading. *J. Literacy Res.* 33 (4), 681–706. doi:10.1080/10862960109548129

Berk, L. E. (2009). *Child Development*. 3rd Canadian ed. Boston, MA: Pearson.

Blythe, M. (2014). Research through Design Fiction: Narrative in Real and Imaginary Abstracts. in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 703–712.

Boeglin-Quintana, B., and Donovan, L. (2013). Storytime Using iPods: Using Technology to Reach All Learners. *Techtrends Tech. Trends* 57 (6), 49–56. doi:10.1007/s11528-013-0701-x

Borgstrom, L. (2011). Developing Story-Time: The Importance of Interactivity in Encouraging Childhood Reading. *Mousaion* 29 (3), 193–208. Available at: https://hdl.handle.net/10520/EJC168980.

Bowlby, J. (1977). The Making and Breaking of Affectional Bonds. *Br. J. Psychiatry* 130 (3), 201–210. doi:10.1192/bjp.130.3.201

Bowman, C. (2012). *Culture of American Families: A National Survey*. Institute for Advanced Studies In Culture. Charlottesville: Institute for Advanced Studies in Culture.

Boyd, D. (2014). *It's Complicated: The Social Lives of Networked Teens*. 1st ed. New Haven: Yale University Press.

Breazeal, C. (2004). Social Interactions in HRI: The Robot View. *IEEE Trans. Syst. Man. Cybern. C.* 34 (2), 181–186. doi:10.1109/tsmcc.2004.826268

Broadbent, E., Stafford, R., and MacDonald, B. (2009). Acceptance of Healthcare Robots for the Older Population: Review and Future Directions. *Int. J. Soc. Robotics*. 1 (4), 319–330. doi:10.1007/s12369-009-0030-6

Bronfenbrenner, U. (1979). *The Ecology of Human Development*. Cambridge, MA: Harvard University Press.

Bruner, J. (1981). The Social Context of Language Acquisition. *Lang. Commun.* 1 (2–3), 155–178. doi:10.1016/0271-5309(81)90010-0

Burns, M. S., Griffin, P., and Snow, C. E. (1999). *Starting Out Right: A Guide to Promoting Children's Reading Success*. Washington, DC: National Academy Press.

Bus, A. G., Belsky, J., van Ijzendoom, M. H., and Crnic, K. (1997). Attachment and Bookreading Patterns: A Study of Mothers, Fathers, and Their Toddlers. *Early Child. Res. Q.* 12 (1), 81–98. doi:10.1016/s0885-2006(97)90044-2

Casillas, M., Bobb, S. C., and Clark, E. V. (2016). Turn-taking, Timing, and Planning in Early Language Acquisition. *J. Child. Lang.* 43 (6), 1310–1337. doi:10.1017/s0305000915000689

Casillas, M., and Frank, M. C. (2017). The Development of Children's Ability to Track and Predict Turn Structure in Conversation. *J. Mem. Lang.* 92, 234–253. doi:10.1016/j.jml.2016.06.013

Chang, S. M., and Sung, H. C. (2013). The Effectiveness of Paro Robot Therapy on Mood of Older Adults: A Systematic Review. *Int. J. Evidence-Based Healthc.* 11 (3), 216. doi:10.1097/01258363-201309000-00018

Cheon, E., and Su, N. M. (2017). Configuring the User: "Robots Have Needs Too" in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (New York, NY: ACM), 191–206.

Chomsky, C. (1972). Stages in Language Development and Reading Exposure. *Harv. Educ. Rev.* 42 (1), 1–33. doi:10.17763/haer.42.1.h78l676h28331480

Chou, M.-J., and Fen, C.-F. (2014). Parent-child Play within Information Technology: A Quest for Quality Family Atmosphere. *Proced. - Soc. Behav. Sci.* 122, 273–282. doi:10.1016/j.sbspro.2014.01.1342

Clark, L. S. (2011). Parental Mediation Theory for the Digital Age. *Commun. Theor.* 21 (4), 323–343. doi:10.1111/j.1468-2885.2011.01391.x

Clarke, J. (2008). All Too Human: Edward Scissorhands. *Screen. Education* (50), 93–98.

Corbin, J., and Strauss, A. (2008). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 3rd ed. Thousand Oaks, CA: Sage. doi:10.4135/9781452230153

Dautenhahn, K., and Werry, I. (2004). Towards Interactive Robots in Autism Therapy. *P&C.* 12 (1), 1–35. doi:10.1075/pc.12.1.03dau

Davis, F. D., Bagozzi, R. P., and Warshaw, P. R. (1989). User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Management Sci.* 35 (8), 982–1003. doi:10.1287/mnsc.35.8.982

De Graaf, M. M. A., Allouch, S. B., and Klamer, T. (2015). Sharing a Life with Harvey: Exploring the Acceptance of and Relationship-Building with a Social Robot. *Comput. Hum. Behav.* 43, 1–14. doi:10.1016/j.chb.2014.10.030

De Graaf, M. M. A., and Ben Allouch, S. (2013). Exploring Influencing Variables for the Acceptance of Social Robots. *Robotics Autonomous Syst.* 61 (12), 1476–1486. doi:10.1016/j.robot.2013.07.007

De Graaf, M. M. A., Ben Allouch, S., and van Dijk, J. A. G. M. (2017). Why Would I Use This in My Home? A Model of Domestic Social Robot Acceptance. *Human-Computer Interaction* 34 (2), 115–173. doi:10.1080/07370024.2017.1312406

De Graaf, M. M. A., and Malle, B. (2019). People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (New York, NY: ACM), 239–248.

Deniz Can, D., and Ginsburg-Block, M. (2016). Parenting Stress and Home-Based Literacy Interactions in Low-Income Preschool Families. *J. Appl. Developmental Psychol.* 46, 51–62. doi:10.1016/j.appdev.2016.07.002

Dunne, A., and Raby, F. (2013). *Speculative Everything: Design, Fiction, and Social Dreaming*. London, United Kingdom: MIT Press.

Eagle, S. (2012). Learning in the Early Years: Social Interactions Around Picturebooks, Puzzles and Digital Technologies. *Comput. Education*. 59 (1), 38–49. doi:10.1016/j.compedu.2011.10.013

Efthymiou, C., and Halvey, M. (2016). Evaluating the Social Acceptability of Voice Based Smartwatch Search. *Asia Inf. Retrieval Symp.* 9994, 267–278. doi:10.1007/978-3-319-48051-0_20

Fabiane, B. (2012). Exploring the Educational Potential of Robotics in Schools: A Systematic Review. *Comput. Education* 58 (3), 978–988. doi:10.1016/j.compedu.2011.10.006

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Evanston, IL: Row, Peterson, and Company.

Fiese, B., and Winter, M. (2009). Family Stories and Rituals. in *Handbook of Family Psychology*. Oxford, UK: Blackwell, 625–636.

Flewitt, R. (2006). Using Video to Investigate Preschool Classroom Interaction: Education Research Assumptions and Methodological Practices. *Vis. Commun.* 5 (1), 25–50. doi:10.1177/1470357206060917

Franklin, D., and Bankston, L. (1999). *Ties that Bind: Building and Maintaining Strong Parent–Child Relationships through Family Traditions and Rituals*. Madison, WI: Wisconsin University.

Frude, N., and Killick, S. (2011). Family Storytelling and the Attachment Relationship. *Psychodynamic Pract.* 17 (4), 441–455. doi:10.1080/14753634.2011.609025

Gates, B. (2007). A Robot in Every Home. *Sci. Am.* 296 (1), 58–65. doi:10.1038/scientificamerican0107-58

Giles, G., and Price, I. R. (2008). Adolescent Computer Use: Approach, Avoidance, and Parental Control. *Aust. J. Psychol.* 60 (2), 63–71. doi:10.1080/00049530701829896

Glaser, B. (1978). *Theoretical Sensitivity: Advances in the Methodology of Grounded Theory*. Mill Valley, CA: Sociology Press.

Gordon, G., Spaulding, S., Kory-Westlund, J. M., Lee, J. J., Plummer, L., Martinez, M., et al. (2016). Affective Personalization of a Social Robot Tutor for Children's Second Language Skills. in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 3951–3957.

Gray, H. M., Gray, K., and Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*. 315 (5812), 619. doi:10.1126/science.1134475

Gray, K., and Wegner, D. M. (2012). Feeling Robots and Human Zombies: Mind Perception and the Uncanny Valley. *Cognition*. 125 (1), 125–130. doi:10.1016/j.cognition.2012.06.007

Gray, K., Young, L., and Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychol. Inq.* 23 (2), 101–124. doi:10.1080/1047840x.2012.651387

Green, P., and Wei-Haas, L. (1985). The Rapid Development of User Interfaces: Experience with the Wizard of Oz Method. *Proc. Hum. Factors Soc. Annu. Meet.* 29 (5), 470–474. doi:10.1177/154193128502900515

Guest, G., Bunce, A., and Johnson, L. (2006). How Many Interviews Are Enough? an Experiment with Data Saturation and Variability. *Field Methods*. 18 (1), 24. doi:10.1177/1525822x05279903

Haight, W. L., Wang, X.-l., Fung, H. H.-t., Williams, K., and Mintz, J. (1999). Universal, Developmental, and Variable Aspects of Young Children's Play: A Cross-Cultural Comparison of Pretending at Home. *Child. Development*. 70 (6), 1477–1488. doi:10.1111/1467-8624.00107

Hales, D. (2013). Design Fictions an Introduction and Provisional Taxonomy. *Digital Creativity*. 24 (1), 1–10. doi:10.1080/14626268.2013.769453

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., and Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Hum. Factors*. 53 (5), 517–527. doi:10.1177/0018720811417254

Harnad, S. (1991). Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem. *Minds and Machines* 1 (1), 43–54. doi:10.1007/BF00360578

Hart, B., and Risley, T. (1999). *The Social World of Children: Learning to Talk*. Baltimore, MD: Brookes Publishing.

H. Ishiguro and F. Dalla Libera in *Geminoid Studies: Science and Technologies for Humanlike Teleoperated Androids* (Singapore: Springer).

Hoffman, G., and Ju, W. (2014). Designing Robots with Movement in Mind. *J. Human–Robot Interaction*. 3 (1), 91–122. doi:10.5898/JHRI.3.1.Hoffman

Horowitz-Kraus, T., and Hutton, J. S. (2018). Brain Connectivity in Children Is Increased by the Time They Spend Reading Books and Decreased by the Length of Exposure to Screen-Based Media. *Acta Paediatr.* 107 (4), 685–693. doi:10.1111/apa.14176

Horst, J. S. (2013). Context and Repetition in Word Learning. *Front. Psychol.* 4, 149. doi:10.3389/fpsyg.2013.00149

Jiow, H. J., Lim, S. S., and Lin, J. (2017). Level up! Refreshing Parental Mediation Theory for Our Digital Media Landscape. *Commun. Theor.* 27 (3), 309–328. doi:10.1111/comt.12109

Kätsyri, J., Förger, K., Mäkäräinen, M., and Takala, T. (2015). A Review of Empirical Evidence on Different Uncanny Valley Hypotheses: Support for Perceptual Mismatch as One Road to the Valley of Eeriness. *Front. Psychol.* 6, 390. doi:10.3389/fpsyg.2015.00390

Kellas, J. K. (2013). *Family Storytelling: Negotiating Identities, Teaching Lessons, and Making Meaning*. London: Routledge.

Kellas, J. K., and Trees, A. R. (2013). Family Stories and Storytelling. in *Handbook of Family Communication*. 2nd ed. Mahwah, NJ: Lawrence Erlbaum.

Kim, S. Y. (1999). The Effects of Storytelling and Pretend Play on Cognitive Processes, Short-Term and Long-Term Narrative Recall. *Child. Study J.* 29 (3), 175–191.

Korn, C. (1998). How Young Children Make Sense of Their Life Stories. *Early Child. Education J.* 25 (4), 223–228. doi:10.1023/a:1025638619861

Kory-Westlund, J. M., and Breazeal, C. (2014). Storytelling with Robots: Learning Companions for Preschool Children's Language Development. in The 23rd IEEE International Symposium on Robot and Human Interactive Communication, 643–648. doi:10.1109/ROMAN.2014.6926325

Kory-Westlund, J. M., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., et al. (2017). Flat vs. Expressive Storytelling: Young Children's Learning and Retention of a Social Robot's Narrative. *Front. Hum. Neurosci.* 11, 295. doi:10.3389/fnhum.2017.00295

Kory-Westlund, J. M., Martinez, M., Archie, M., Das, M., and Breazeal, C. (2016). A Study to Measure the Effect of Framing a Robot as a Social Agent or as a Machine on Children's Social Behavior. in Proceedings of the 11th ACM/IEEE International Conference on Human–Robot Interaction (New York: ACM), 459–460. doi:10.1109/HRI.2016.7451805

Kory-Westlund, J. M., Park, H. W., Williams, R., and Breazeal, C. (2018). Measuring Young Children's Long-Term Relationships with Social Robots. in Proceedings of the 17th ACM Conference on Interaction Design and Children (New York: ACM), 207–218. doi:10.1145/3202185.3202732

Kossewska, J., and Kłosowska, J. (2020). Acceptance of Robot-Mediated Teaching and Therapy for Children with Atypical Development by Polish Professionals. *J. Policy Pract. Intellect. Disabi.*. 17 (1), 21–30. doi:10.1111/jppi.12296

Langellier, K., and Family, E. P. (2006). "Narrative Performance Theory: Telling Stories, Doing Family," in *Engaging Theories in Family Communication: Multiple Perspectives*. Editors D. O. Braithwaite and L. A. Baxter (Thousand Oaks, CA: Sage), 99–114.

Laughlin, L. (2013). *Who's Minding the Kids? Child Care Arrangements: Spring 2011*. Washington, DC: US Census Bureau.

Lauricella, A. R., Wartella, E., and Rideout, V. J. (2015). Young Children's Screen Time: The Complex Role of Parent and Child Factors. *J. Appl. Developmental Psychol.* 36, 11–17. doi:10.1016/j.appdev.2014.12.001

Levin, D. E., and Rosenquest, B. (2001). The Increasing Role of Electronic Toys in the Lives of Infants and Toddlers: Should We Be Concerned? *Contemp. Issues Early Child.* 2 (2), 242–247. doi:10.2304/ciec.2001.2.2.9

Lin, C. H., Liu, E. Z. F., and Huang, Y. Y. (2012). Exploring Parents' Perceptions towards Educational Robots: Gender and Socio-Economic Differences. *Br. J. Educ. Technology* 43 (1), E31–E34. doi:10.1111/j.1467-8535.2011.01258.x

Lindblom, J., and Andreasson, R. (2016). Current Challenges for UX Evaluation of Human-Robot Interaction. in *Advances in Intelligent Systems and Computing*. Cham, Switzerland: Springer, 267–277. doi:10.1007/978-3-319-41697-7_24

Lindley, J., and Coulton, P. (2015). Back to the Future: 10 Years of Design Fiction. in Proceedings of the 2015 British HCI Conference, 210–211. doi:10.1145/2783446.2783592

Livingstone, S., and Helsper, E. J. (2008). Parental Mediation of Children's Internet Use. *J. Broadcasting Electron. Media* 52 (4), 581–599. doi:10.1080/08838150802437396

Lynd, R. S., and Lynd, H. M. (1929). *Middletown, a Study in Contemporary American Culture*. New York, NY: Harcourt, Brace.

MacDorman, K. F., and Cowley, S. J. (2006). "Long-term Relationships as a Benchmark for Robot Personhood," in Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication, September 6–9, 2006 (Hatfield, UK: University of Hertfordshire), 378–383.

MacDorman, K. F., and Ishiguro, H. (2006). The Uncanny Advantage of Using Androids in Cognitive and Social Science Research. *Interaction Studies*. 7 (3), 297–337. doi:10.1075/is.7.3.03mac

MacDorman, K. F., Vasudevan, S. K., and Ho, C.-C. (2009). Does Japan Really Have Robot Mania? Comparing Attitudes by Implicit and Explicit Measures. *AI Soc.* 23 (4), 485–510. doi:10.1007/s00146-008-0181-2

Malle, B. F., and Scheutz, M. (2018). *Learning How to Behave. Handbuch Maschinenethik*. Wiesbaden, Germany: Springer, 1–24. doi:10.1007/978-3-658-17484-2_17-1

Malpass, M. (2013). Between Wit and Reason: Defining Associative, Speculative, and Critical Design in Practice. *Des. Cult.* 5 (3), 333–356. doi:10.2752/175470813x13705953612200

Mascheroni, G., Ponte, C., and Jorge, A. (2018). *Digital Parenting: The Challenges for Families in the Digital Age*. Sweden: Göteborg.

McFarlin, L. A., Buffardi, K. J., and Schumacher, R. M. (2007). Usability Impact on Effectiveness of Parental Controls. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 51 (17), 1039–1043. doi:10.1177/154193120705101709

Miklósi, Á., Korondi, P., Matellán, V., and Gácsi, M. (2017). Ethorobotics: A New Approach to Human–Robot Relationship. *Front. Psychol.* 8, 958. doi:10.3389/fpsyg.2017.00958

Montero, C. S., Alexander, J., Marshall, M. T., and Subramanian, S. (2010). Would You Do that? Understanding Social Acceptance of Gestural Interfaces. in Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, 275–278. doi:10.1145/1851600.1851647

Mori, M. (2012). The Uncanny Valley (K. F. MacDorman & Norri Kageki, Trans.). *IEEE Robotics & Automation*. 19 (2), 98–100. doi:10.1109/MRA.2012.2192811

Nielsen, J. (1993). *Usability Engineering*. San Diego, CA: Academic Press.

Nikken, P., and Jansz, J. (2014). Developing Scales to Measure Parental Mediation of Young Children's Internet Use. *Learn. Media Technology* 39 (2), 250–266. doi:10.1080/17439884.2013.782038

Nisbett, R. E., and Wilson, T. D. (1977). The Halo Effect: Evidence for Unconscious Alteration of Judgments. *J. Personal. Soc. Psychol.* 35 (4), 250–256. doi:10.1037/0022-3514.35.4.250

Otway, H. J., and Von Winterfeldt, D. (1982). Beyond Acceptable Risk: On the Social Acceptability of Technologies. *Policy Sci* 14 (3), 247–256. doi:10.1007/bf00136399

Paetzel, M., Perugia, G., and Castellano, G. (2020). The Persistence of First Impressions: The Effect of Repeated Interactions on the Perception of a Social Robot. in Proceedings of the ACM/IEEE International Conference on Human–Robot Interaction (New York: ACM), 73–82. doi:10.1145/3319502.3374786

Peltonen, P., Kurvinen, E., Salovaara, A., Jacucci, G., Ilmonen, T., Evans, J., et al. (2008). It's Mine, Don't Touch! Interactions at a Large Multi-Touch Display in a City Centre. in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York: ACM), 1285–1294. doi:10.1145/1357054.1357255

Plowman, L. (2015). Researching Young Children's Everyday Uses of Technology in the Family Home. *Interact. Comput.* 27 (1), 36–46. doi:10.1093/iwc/iwu031

Plowman, L. (2004). "Hey, Hey, Hey! It's Time to Play": Children's Interactions with Smart Toys. in *Toys, Games, and Media*. Mahwah, NJ: L. Erlbaum Associates, 207–223.

Porteous, J., Charles, F., Smith, C., Cavazza, M., Mouw, J., and Van Den Broek, P. (2017). Using Virtual Narratives to Explore Children's Story Understanding. *Int. Conf. Autonomous Agents Multiagent Syst.*, 773–781.

Quinton, A. M. (1955). The Problem of Perception. *Mind*. LXIV (253), 28–51. doi:10.1093/mind/lxiv.253.28

Reeves, B., and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. New York: Cambridge University Press.

Robertson, J. (2007). Robo Sapiens Japanicus: Humanoid Robots and the Posthuman Family. *Crit. Asian Stud.* 39 (3), 369–398. doi:10.1080/14672710701527378

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., et al. (2018). Beyond the 30-million-word Gap: Children's Conversational

Exposure Is Associated with Language-Related Brain Function. *Psychol. Sci.* 29 (5), 700–710. doi:10.1177/0956797617742725

Šabanović, S. (2010). Robots in Society, Society in Robots. *Int. J. Soc. Robotics*. 2 (4), 439–450. doi:10.1007/s12369-010-0066-7

Savage, B. M., Lujan, H. L., Thipparthi, R. R., and DiCarlo, S. E. (2017). Humor, Laughter, Learning, and Health! A Brief Review. *Adv. Physiol. Education*. 41 (3), 341–347. doi:10.1152/advan.00030.2017

Saygin, A. P., Cicekli, I., and Akman, V. (2000). Turing Test: 50 Years Later. *Minds and Machines*. 10 (4), 463–518. doi:10.1023/A:1011288000451

Scassellati, B., Henny Admoni, H., and Matarić, M. (2012). Robots for Use in Autism Research. *Annu. Rev. Biomed. Eng.* 14, 275–294. doi:10.1146/annurev-bioeng-071811-150036

Schneider, P., Weisner, T. S., and Gallimore, R. (1997). Family and Child Factors Contributing to "Successful Storytime Interactions" with Children with Developmental Delays. *Developmental Disabilities Bull.* 25 (2), 32–49.

Schulte, B. F., Marshall, P., and Cox, A. L. (2016). Homes for Life: A Design Fiction Probe. *Proc. Ninth Nordic Conf. Human–Computer Interaction*, 1–10. doi:10.1145/2971485.2993925

Searle, J. R. (1980). Minds, Brains, and Programs. *Behav. Brain Sci.* 3 (3), 417–424. doi:10.1017/s0140525x00005756

Segal, H. P. (1986). "The Technological Utopians," in *Imagining Tomorrow: History, Technology and the American Future*. Editor J. J. Corn (Cambridge, MA: MIT Press), 119–136.

Shanahan, T., and Lonigan, C. (2013). *Early Childhood Literacy: The National Early Literacy Panel and Beyond*. Baltimore, MD: Brookes Publishing.

Sharkey, N., and Sharkey, A. (2010). The Crying Shame of Robot Nannies. *Interaction Studies*. 11 (2), 161–190. doi:10.1075/is.11.2.01sha

Shibata, T., and Wada, K. (2011). Robot Therapy: A New Approach for Mental Healthcare of the Elderly – A Mini-Review. *Gerontology*. 57 (4), 378–386. doi:10.1159/000319015

Silvern, S. (1985). Parent Involvement and Reading Achievement: A Review of Research and Implications for Practice. *Child. Education*. 62 (1), 44–50. doi:10.1080/00094056.1985.10520221

Soundy, C. S. (1993). Let the Story Begin! Open the Box and Set Out the Props. *Child. Education*. 69 (3), 146–149. doi:10.1080/00094056.1993.10520915

Speaker, K. M. (2000). The Art of Storytelling: A Collegiate Connection to Professional Development Schools. *Education*. 121 (1), 184

Stanley, N., and Dillingham, B. (2009). *Performance Literacy through Storytelling*. Gainessville: Maupin House Publishing.

Sullivan, R., Perry, R., Sloan, A., Kleinhaus, K., and Burtchen, N. (2011). Infant Bonding and Attachment to the Caregiver: Insights from Basic and Clinical Science. *Clin. Perinatology*. 38 (4), 643–655. doi:10.1016/j.clp.2011.08.011

Sulzby, E. (1985). Children's Emergent Reading of Favorite Storybooks: A Developmental Study. *Selections*. 20 (4), 458–481. doi:10.1598/rrq.20.4.4

Tazhigaliyeva, N., Diyas, Y., Brakk, D., Aimambetov, Y., and Sandygulova, A. (2016). Learning with or from the Robot: Exploring Robot Roles in Educational Context with Children. *Lecture Notes Computer Sci.* 9979, 650–659. doi:10.1007/978-3-319-47437-3_64

Thakur, V. (2016). People's Perception of Smart Home Devices in Their Homes & the Factors Which Influence This Perception. Master's thesis, Utica (NY): State University of New York. Available at: http://hdl.handle.net/1951/68894.

Thrun, S. (2004). Toward a Framework for Human-Robot Interaction. *Human-comp. Interaction*. 19 (1), 9–24. doi:10.1207/s15327051hci1901&2_2

Toh, L. P. E., Causo, A., Tzuo, P. W., Chen, I. M., and Yeo, S. H. (2016). A Review on the Use of Robots in Education and Young Children. *J. Educ. Technology Soc.* 19 (2), 148–163. Available at: https://www.jstor.org/stable/10.2307/jeductechsoci.19.2.148.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*. LIX (236), 433–460. doi:10.1093/mind/lix.236.433

Turkle, S. (2007). "Authenticity in the Age of Digital Companions," in *Psychological Benchmarks of Human–Robot Interaction*. Editors P. H. KahnJr. and K. F. MacDorman, 8, 501–517. doi:10.1075/is.8.3.11tur

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Q.* 27 (3), 425–478. doi:10.2307/30036540

Wallkötter, S., Stower, R., Kappas, A., and Castellano, G. (2020). A Robot by Any Other Frame: Framing and Behaviour Influence Mind Perception in Virtual but Not Real-World Environments. in Proceedings of the 2020 ACM/IEEE International Conference on Human–Robot Interaction (New York, NY: ACM), 609–618. doi:10.1145/3319502.3374800

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (New York, NY: ACM), 1–15. doi:10.1145/3290605.3300831

Wang, S., Lilienfeld, S. O., and Rochat, P. (2015). The Uncanny Valley: Existence and Explanations. *Rev. Gen. Psychol.* 19 (4), 393–407. doi:10.1037/gpr0000056

Wells, G. (1986). *The Meaning Makers: Children Learning Language and Using Language to Learn*. Portsmouth, NH: Heinemann.

Whitehead, B. D. (1991). The Family in an Unfriendly Culture. *Fam. Aff.* 3 (1–2), 1–6.

Willemse, C. J. A. M., Toet, A., and van Erp, J. B. F. (2017). Affective and Behavioral Responses to Robot-Initiated Social Touch: Toward Understanding the Opportunities and Limitations of Physical Contact in Human–Robot Interaction. *Front. ICT*. 4, 12. doi:10.3389/fict.2017.00012

Yang, G.-Z., Nelson, B. J., Murphy, R. R., Choset, H., Christensen, H., Collins, S. H., et al. (2020). Combating COVID-19-The Role of Robotics in Managing Public Health and Infectious Diseases. *Sci. Robot.* 5 (40), eabb5589. doi:10.1126/scirobotics.abb5589

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

The handling Editor declared a past co-authorship with one of the authors (SS).

frontiers in Robotics and AI

ORIGINAL RESEARCH
published: 31 May 2021
doi: 10.3389/frobt.2021.676123

Do Shy Preschoolers Interact Differently When Learning Language With a Social Robot? An Analysis of Interactional Behavior and Word Learning

Nils F. Tolksdorf, Franziska E. Viertel and Katharina J. Rohlfing*

Faculty of Arts and Humanities, Psycholinguistics, Paderborn University, Paderborn, Germany

Temperamental traits can decisively influence how children enter into social interaction with their environment. Yet, in the field of child–robot interaction, little is known about how individual differences such as shyness impact on how children interact with social robots in educational settings. The present study systematically assessed the temperament of 28 preschool children aged 4–5 years in order to investigate the role of shyness within a dyadic child–robot interaction. Over the course of four consecutive sessions, we observed how shy compared to nonshy children interacted with a social robot during a word-learning educational setting and how shyness influenced children's learning outcomes. Overall, results suggested that shy children not only interacted differently with a robot compared to nonshy children, but also changed their behavior over the course of the sessions. Critically, shy children interacted less expressively with the robot in general. With regard to children's language learning outcomes, shy children scored lower on an initial posttest, but were able to close this gap on a later test, resulting in all children retrieving the learned words on a similar level. When intertest learning gain was considered, regression analyses even confirmed a positive predictive role of shyness on language learning gains. Findings are discussed with regard to the role of shyness in educational settings with social robots and the implications for future interaction design.

Keywords: child–robot interaction, temperament, shyness, early childhood education, social robot, personality and behavior, word learning, individual differences

Edited by:
Tony Belpaeme,
Ghent University, Belgium

Reviewed by:
Samuel Spaulding,
Massachusetts Institute of Technology, United States
Sofia Serholt,
University of Gothenburg, Sweden

****Correspondence:***
Nils F. Tolksdorf
nils.tolksdorf@upb.de

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: *04 March 2021*
Accepted: *23 April 2021*
Published: *31 May 2021*

Citation:
Tolksdorf NF, Viertel FE and Rohlfing KJ (2021) Do Shy Preschoolers Interact Differently When Learning Language With a Social Robot? An Analysis of Interactional Behavior and Word Learning. Front. Robot. AI 8:676123. doi: 10.3389/frobt.2021.676123

INTRODUCTION

Recent years have seen a substantial growth in the applicability of social robots in educational learning environments with young learners. Examples are therapeutic settings (Boccanfuso et al., 2017; Cao et al., 2019), science learning (Causo et al., 2016; Park et al., 2017; Ioannou and Makridou, 2018), or language learning (van den Berghe et al., 2019; Vogt et al., 2019). More specifically, with the benefits of an embodied agent (Belpaeme et al., 2018), social robots offer versatile possibilities to engage children systematically in social interaction. Indeed, current research suggests that children accept social robots as trustworthy social actors from whom they can obtain reliable information (Breazeal et al., 2016; Vollmer et al., 2018; Oranç and Küntay, 2020). However, whereas these findings consider the demonstrated behavior displayed by children on average, individual differences in child–robot interactions have received little attention. Nonetheless, the way children enter into social interaction with their environment is mediated crucially by their individual temperament—specifically, by their shyness (Coplan and Evans, 2009). In fact, past research has demonstrated that a child's shyness significantly influences her or his social behavior in familiar and unfamiliar contexts (Reddy, 2000; Evans, 2001; Coplan and Weeks, 2009; Crozier and Badawood, 2009; Feinberg et al., 2012; Colonnesi et al., 2014; Smith Watts et al., 2014). Additionally, shyness has a substantial impact on children's performance in educational settings, insofar as shy children can either struggle to demonstrate their abilities in such situations or implicitly reduce their learning opportunities because they avoid social interaction (Evans, 2001; Spere et al., 2004; Smith Watts et al., 2014). Although some work has focused on how individual personality traits of adults affect interaction with a robot (Walters et al., 2005; Salter et al., 2006; Salam et al., 2017), research accounting for children's behavior lags behind. Thus, our aim is to raise awareness in the area of child–robot interaction about how children's personality traits such as shyness are reflected in their behavior when interacting with a social robot. Specifically, we aim to understand how the behavior of shy children might develop over a long-term interaction and influence learning gains with a social robot.

Extensive past research underlines the prevalence of shyness, indicating that up to 90% of the population experience shyness at some point in their lives with about 15% of individuals displaying a shyness that emerges in early development and remains stable across contexts (Zimbardo et al., 1975; Kagan, 1994; Schmidt et al., 2020). Shyness in children can be conceptualized as an increased and persistent behavioral inhibition in unfamiliar social situations or during perceived social evaluation that can result in withdrawal from interaction (Putnam et al., 2006; Rubin et al., 2009; Barker et al., 2014; see Schmidt and Buss, 2010, for a review). In developmental research, the effects of shyness are well documented, showing that the behavior of shy children toward their environment is reflected in both their verbal (Coplan and Weeks, 2009; Crozier and Badawood, 2009; Smith Watts et al., 2014) and nonverbal behavior (Reddy, 2000; Evans, 2001; Putnam et al., 2006; Feinberg et al., 2012; Colonnesi et al., 2014). For example, shy children are less talkative in familiar and unfamiliar contexts (Asendorpf and Meier, 1993; Evans, 1996; Crozier and Badawood, 2009) and display shyness through their facial expressions (e.g., by showing coy smiles) or their gaze behavior (e.g., duration of eye contact or gaze and head aversion; see Reddy, 2005; Colonnesi et al., 2014; Colonnesi et al., 2017). Importantly, recent research emphasizes that such expressions of shyness provide a positive and socially adaptive function for shy children within an interaction that enables them to regulate their emotions in unknown situations while also increasing prosociality and trust (Reddy, 2005; Colonnesi et al., 2013; Colonnesi et al., 2020).

Beyond the fact that shyness is reflected in children's behavior toward an interaction partner, shyness also has an effect on the measured learning performance in educational contexts and particularly in the domain of language learning. Spere et al. (2004) and Spere and Evans (2009) have demonstrated that temperamentally shy preschoolers perform more poorly on both expressive and receptive vocabulary tests compared to nonshy children. Similarly, Hilton and Westermann (2017), recently investigated shy children's retention abilities for learned word–object mappings and compared their learning outcomes with those of nonshy children. They looked at long-term word learning processes when the children were engaged in a word learning task: After a 5 min break, shy children did not retain any novel word they had formed during the learning situation; less shy children, in contrast, were able to retain the trained words. In line with work indicating that shy children are less likely to take risks in situations that are unknown or in which they are being evaluated (Addison and Schmidt, 1999; Levin and Hart, 2003), the authors suggested that shy children rely less on a guess in their responses, and that this might be detrimental in a situation in which they have to retrieve a novel word with ambiguous referents (Hilton and Westermann, 2017). Additionally, they argued that shyness may affect not only performance during an evaluation situation but also the immediate learning process—that is, due to shy children's aversion to the unfamiliar, they might be less inclined to use novelty as a cue to the appropriate referent of a novel label when familiar/competitor objects are present. However, in this respect, it has been suggested that these differences in learning achievements could be context-dependent and not genuine, and that they should disappear under conditions that minimize anxiety or fear of evaluation (Spere et al., 2004; Hilton and Westermann, 2017). Along these lines, it should be borne in mind that conducting an experiment typically represents a fundamentally unfamiliar social situation for a child through, for instance, being exposed to unfamiliar people and unfamiliar settings. Thus, and against the background of evidence that shy children tend to show difficulties in experimental tasks (Crozier and Hostettler, 2003), it can be argued that the unfamiliar contextual environment during the testing situation might have influenced the recall abilities of shy children. Furthermore, Hilton and Westermann (2017), did not assess children's general linguistic skills, although research has shown that existing linguistic knowledge should be considered when measuring word-learning processes (Stelmachowicz et al., 2004; McMurray et al., 2012). Therefore, investigations that include children's linguistic abilities, are extended over a longer period of time, and are not limited to a single occasion could increase familiarity with the contextual environment and shed a more nuanced light on shy children's learning outcomes in comparison to their nonshy peers. In sum, the currently available evidence shows that shy children's behavior and learning is highly sensitive to contextual changes that may be particularly pronounced in social interactions such as in unfamiliar face-to-face encounters or during testing situations (Spere et al., 2004; Putnam et al., 2006; Matsuda et al., 2013).

Turning to the area of child–robot interaction, only a few attempts have been made to explore shy children's behavior during interaction with a social robot, although possible implications of shyness for child–robot interaction in educational contexts are acknowledged (Baxter and Belpaeme, 2016; Baxter et al., 2017). Instead, most studies dealing with personality traits in this field examine approaches to provide the robotic system with the ability to automatically estimate the personality of a child based on predefined behavioral characteristics (Abe et al., 2014; Abe et al., 2017; Schodde et al., 2017; Abe et al., 2020; Sano et al., 2020) or to equip the robot itself with certain personality traits (Fischer et al., 2019; Calvo-Barajas et al., 2020). One of the few studies to address shyness in child–robot interaction investigated preschooler's perceptions during a free-play situation in a kindergarten setting (Abe et al., 2014). In this study, parents evaluated their children's shyness on a 5-point scale before estimating the social relationship between their child and the robot after a single "one-off" interaction. The study found that shyness clearly affected the relationship with the robot; and, according to the parents, one third of the shy children lacked a friendly relationship with the robot, whereas almost all nonshy children were friendly with the robot (Abe et al., 2014). Although examining parental reports about an experienced child–robot interaction is a proven methodological approach to assess and contextualize a child's behavior (Tolksdorf and Rohlfing, 2020), how shy children actually behave within a child–robot interaction setting remains an open question. Additionally, given that familiarity with a situation strongly influences shyness (Rimm-Kaufman and Kagan, 2005), there is a general lack of a perspective that would include the long-term effects over multiple interactions.

Other studies that consider the effects of children's individual personality traits are based on more implicit findings: Vogt et al. (2019) reported that in an educational child–robot interaction, certain preschoolers dropped out of the entire interaction due to shyness, or they needed additional support from the caregiver to successfully interact with the robot because they were reluctant or anxious even after an initial introduction to it. Shiomi et al. (2016) made a similar observation, showing that in a free-play situation, almost one quarter of the children hesitated to interact with the robot or avoided interaction entirely. The authors explained this rejection behavior as a result of the inhibition of these children (Shiomi et al., 2016). However, because they did not assess the children's temperament, it is not clear whether their behavior can be linked to their shyness.

In a more recent work, Tolksdorf et al. (2020b) were the first to investigate which expressions of shyness are displayed by a child during an educational child–robot interaction. Their pilot study measured preschool children's shyness with a standardized and validated questionnaire (Zentner, 2011) and analyzed children's expressions of shyness toward the social robot across three sessions. In interaction with the robot, shy children not only behaved differently, but also changed their behavior over the course of the sessions: Although they showed significantly less positive behavior in the first interaction compared to the nonshy children, these differences disappeared in subsequent sessions. The authors argued that this could be explained in terms of increasing familiarity with the novel interaction partner, and that shy children might be able to overcome their reluctance to interact with a robotic system. In fact, these results are in line with work suggesting that young children react with uncertainty when facing a robot for the first time, and that they rely on an adequate introduction by a familiar caregiver to establish a beneficial learning environment (Rohlfing et al., 2020).

Taken together, these few studies indicate clearly that individual differences in the behavioral style toward a social robot exist in children, and they can be related to children's shyness. However, any generalization across previous reports on the relation between shyness and children's interaction behavior toward a robot is difficult because of the differences in the precise operationalization and assessment of shyness. Importantly, considering that social robots are being evaluated increasingly as learning partners, we do not know how far temperamental characteristics such as shyness might influence a child's learning gain. Furthermore, although earlier studies evidenced that some behavioral effects in shy children's behavior disappear when they are given the opportunity to familiarize themselves with the situation (Evans, 2001; Rimm-Kaufman and Kagan, 2005; Arbeau et al., 2010), the literature lacks a perspective focusing on how the behavior of shy children develops during a long-term child–robot interaction over multiple points in time and including multiple exposures to a test situation. Therefore, following up on our previous work, the current study aimed to address this research gap: We investigated the impact of shyness during an educational child–robot interaction by systematically examining children's learning performance as well as their interaction behavior in terms of shyness markers and their signals of pleasure and distress toward the robot over the course of a long-term study on language learning.

In the present study, preschool-age children took part in a child–robot interaction over three consecutive learning situations followed by two test situations within a time period of two weeks. In line with recent accounts arguing that young children react with uncertainty when initially encountering a robot (Vogt et al., 2017; Rohlfing et al., 2020), we assumed that all children would interact with a certain reluctance at the beginning of the long-term interaction because they were faced with a novel and unfamiliar situation. Our main research interest was to explore how shy children's behavior and language learning would develop with increasing familiarity during the sessions. We formulated the following hypotheses:

1. **(H1)** Based on the aforementioned work demonstrating shy children's behavioral inhibition in unfamiliar social situations (Evans, 1996; Reddy, 2000; Putnam et al., 2006; Feinberg et al., 2012; Abe et al., 2014; Colonnesi et al., 2014; Vogt et al., 2019; Tolksdorf et al., 2020b), we expected that shy children would show less positive reactions (H1a) and more negative reactions (H1b) than nonshy children in both the learning situations and the test situations with the social robot just like they would be expected to do in human–human face-to-face interactions.
2. **(H2)** We also expected that negative reactions would decrease and positive reactions would increase with the repetition of a situation, especially among the shy children—both when there was a repetition of the learning situation (H2a) as well as when the test situation was repeated (H2b). This was because prior research has shown that the repetition of an interaction

leads to an environment that becomes more predictable for a child while also increasing familiarity with the situation (Bruner, 1983; Rohlfing et al., 2016), and this might result in more positive expressions and minimize children's social discomfort during the interaction (Colonnesi et al., 2014).

3. **(H3)** With regard to children's language learning, we hypothesized that shy children would display lower learning achievements compared to nonshy children. This hypothesis is consistent with work indicating that shy children tend to perform on a lower level when their linguistic knowledge is tested in unfamiliar social situations (Spere et al., 2004; Hilton and Westermann, 2017). However, if growing familiarity with the testing environment allows shy children to feel more comfortable (Hilton et al., 2019), we expected that in the second test situation, the gap between shy and the nonshy children in displayed learning outcomes would decrease or disappear.

MATERIALS AND METHODS

Participants

Thirty preschool children participated in the study. The data from two children were excluded because they did not participate in all sessions. This left 28 children (11 female, 17 male) aged 4.00–5.83 years (mean age = 4.98, *SD* = 0.48) for the final analysis. The children and their parents came from relatively high socioeconomic status backgrounds and were recruited from the wider Paderborn region (North Rhine-Westphalia, Germany). Recruitment was conducted in local kindergartens and libraries or through newspapers and our database of families willing to participate in our research studies. In addition, we assessed the level of parental education. None of the children or the parents had ever seen the robot before the experiment. Prior to their children's participation, parents provided written consent and filled out a questionnaire on their child's general and language development. Criteria for inclusion in the sample were: a) age between 4 and 6 years; b) normal general development such as normal sensory and cognitive skills; and 3) no developmental language delays. Additionally, detailed information regarding children's language skills was obtained by measuring their receptive language abilities with a subtest of the standardized SETK 3–5 (Grimm et al., 2001) and their expressive vocabulary with the AWST-R (Kiese-Himmel, 2005). Parents were present during all interactions, but did not participate actively in the interaction. Children also provided verbal assent prior to taking part in the interaction, and the interaction could be discontinued at any time at no disadvantage to the child. Moreover, each child received stickers and a toy to thank them for their participation.

Experimental Procedure

The children and their parents were invited to visit our laboratory at Paderborn University for four sessions within a period of two weeks. Each session lasted around 20–35 min, and all sessions were recorded on video. Each participating child was accompanied by one parent. **Figure 1** displays the seating arrangement. The experimenter operated the robot, avoiding any interaction with either parent or child. In addition, parents were instructed to avoid talking to their child during the experimental part of the child's interaction with the social robot.

Based on previous work and ethical considerations (Vogt et al., 2019; Tolksdorf et al., 2020a; Tolksdorf and Mertens, 2020), we conducted a warm-up phase with each child and her or his parent before the learning situation with the robot. This introduced the robot to the children in a comfortable way with their caregiver as an available resource (Manner et al., 2018; Tolksdorf et al., 2020a) and reduced the novelty effect (Kanero et al., 2018). During the warm-up, the experimenter first introduced the robot in a powered-off state to the child and parent and explained its functions. For example, it was explained that the robot can talk and move with the help of small motors, because a pilot study had shown that some children were surprised when they heard that the robot's movements were loud. In a second step, children and parents were further familiarized with the capabilities of the robot: The robot introduced itself and performed a short game by imitating animal movements and asking the child, the parent, and the experimenter to repeat the movements. Although the experimenter structured the situation and was the main interaction partner for the child, parental involvement was considered as an important element during the warm-up phase, because prior work has demonstrated that young children may rely on the emotions with which their familiar caregiver interprets the ongoing situation, especially during first encounters with a social robot (Rohlfing et al., 2020; Tolksdorf et al., 2021). After the game was completed, the robot said goodbye for the moment and announced that it had prepared a story that it wanted to share with the child. Subsequently, the experiment started and the script for the first learning situation was launched.

When designing the learning situation, we were guided by theoretical concepts of learning postulating that interaction partners jointly co-construct the communicative situation in a goal-oriented way (Rohlfing et al., 2016; Rohlfing et al., 2019). Therefore, we chose a setting that included activities with which preschoolers are familiar. Specifically, the robot told the child a story that had been created to frame the learning situation. The story contained the plot of the robot's trip to Paderborn University and the things it had seen on its journey. This narrative served as the context in which the children encountered six novel words (color adjectives) during the interaction. The referents of the novel words were presented as pictures hanging on the wall. They were covered by a small cloth, and the robot asked the child to uncover them one by one over the course of the interaction (see **Figure 2A**). This context was also chosen because past work has shown that the context of a story is particularly conducive to children's learning of new words (Horst, 2013; Nachtigäller et al., 2013).

Furthermore, in order to render the robot's interaction behavior child-oriented and to fulfill the important role of multimodal joint activities (Rohlfing et al., 2019; Tolksdorf and Mertens, 2020), the robot also performed a number of

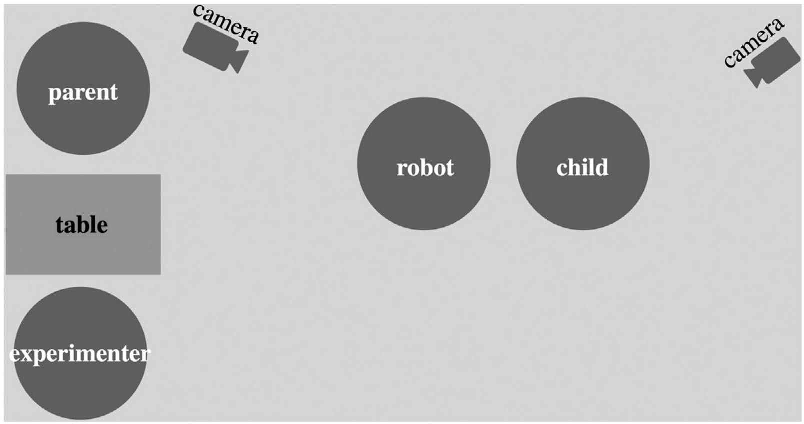


FIGURE 1 | Setup of the study.

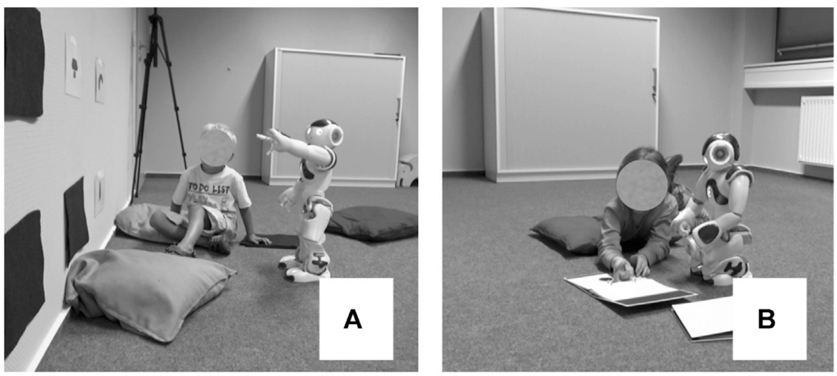


FIGURE 2 | The learning situation **(A)** and test situation **(B)**.

actions such as accompanying the novel words with pointing gestures to coordinate the child's attention and establish a shared reference. In the same way, the robot coordinated its gaze between the child and the referents of the target words. Additionally, after naming the first four target words, the robot also walked with the child to the two remaining target referents in order to make the situation more natural and to take advantage of the physical presence of the robot (van den Berghe et al., 2019). Once the robot had finished the story, it thanked the child and said goodbye.

Subsequently, in the second and third sessions, a similar learning situation with the robot took place again in which the robot shared the story, and the children were exposed to the novel words.

Following the third learning situation in the third session and a 5 min break, children's learning achievements were assessed by testing their ability to generalize the acquired knowledge. In this generalization task, comprehension was defined as the child's ability to extend or transfer the target words to new objects, and whether they were able to transfer the learned pattern of word formation to new colors when presented with separated units of a compound. This method ensured assessment of whether the children were able to transfer their knowledge to other objects and whether their knowledge was stable. We used a routinized activity for children and embedded the test procedure within a shared picture book reading situation (Grimminger and Rohlfing, 2017). In this test, the child was asked to turn the pages while the robot talked about the pictures with the child and elicited the trained words (see **Figure 2B**).

Two to three days later, the last, fourth session was conducted, in which a delayed test using the same procedure was administered to assess the children's knowledge again.

Stimuli

Our study used the Nao robot from Softbank Robotics. This is a small, toy-like, humanoid robot used widely in child–robot interaction studies (Belpaeme et al., 2018). It is 58 cm high with 25 degrees of freedom. Teleoperation was employed to enable the robot to act contingently (Kennedy et al., 2017). We implemented the behaviors in the NAO robot by using *Choregraphe* and used the integrated text-to-speech production of the robot with German language enabled and speech reduced

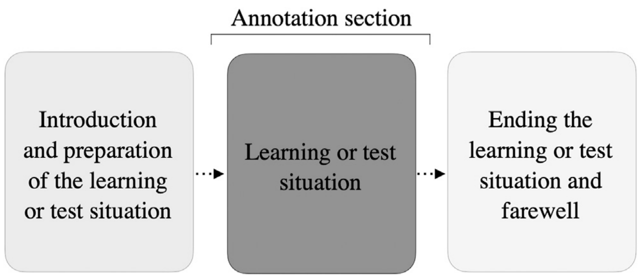


FIGURE 3 | Sequence of the interaction during sessions and annotated section.

to 85% speed to achieve a more natural pronunciation. The target words imparted by the robot were spoken at a speed of 75% in order to emphasize them verbally. The target items consisted of six morphologically complex words (noun–adjective compounds such as "quince yellow [*quittengelb*]") that represented different colors as features of different objects. Each item was presented on a picture measuring 14.8 × 21.0 cm.

Coding of Children's Behavior

Because we were interested in children's behavior during all the learning and testing situations with the social robot, we followed Colonnesi et al. (2013) and Colonnesi et al. (2014) and coded positive as well as negative reactions with and without signs of aversion such as body or head aversion. Positive behaviors of the children were measured by, for example, smiling identified by raising the corners of the lips, constriction of the eyes, raising of the cheeks, or opening of the mouth. Negative behaviors were frowning or sad facial expressions. When these behaviors were accompanied by an aversion (of gaze, head, and body), they fell into the category of positive or negative expressions of shyness. We measured this behavior across the time period of the interaction during which the robot a) shared the story and taught the new words (learning situation T 1–3) and 2) tested the child within the shared picture book situation (Test 1 and Test 2), (see **Figure 3**).

We decided to analyze this sequence because, at this stage, all children had already achieved a certain familiarity with the new interaction partner, and our focus was on investigating development across different sessions. Additionally, the analyzed sequence represented the main part of the interaction, whereas a welcome or farewell situation would reflect a different social situation with its own contextually appropriate social behaviors (Vaughn and La Greca, 1992). Examining this sequence is particularly relevant, because it provides an opportunity to understand how shy children interact during a learning situation with a social robot. Because the duration of the interactions varied slightly between children, children's behavior was expressed in proportion per minute. To evaluate coding reliability, two coders independently coded a random subset of 15% of the data. We used Cohen's Kappa to measure the agreement between the coders for positive and negative reactions, expressions of shyness, and children's aversion. The mean Kappa values were between 0.88 and 0.94, indicating a high level of internal consistency.

Assessment of Naming Performance

Following recent methodological accounts (Rohlfing and Grimminger, 2019), we chose to assess children's word-learning performance in detail on different linguistic layers rather than in a binary way (e.g., only correct or incorrect). To provide a measure of word learning, we created a composite score by averaging each of the children's naming performances in percentages on a phonological, morphosyntactic, and pragmatic–semantic level. Thus, on a phonological level, we calculated the proportion of correctly produced syllables of a target word. On a morphosyntactic level, and independent of the semantic meaning, we assessed the sophistication with which the children constructed a noun–adjective compound and distinguished between no compounding, partial compounding, and fully correct compounding. Finally, we evaluated each child's response on a pragmatic–semantic level, differentiating between no response, a semantically adequate response, a partial retrieval of the target word, and a fully correct retrieval. The maximum composite score that could be achieved in each test session was 3, which would reflect a fully accurate performance on each of the three linguistic levels.

Assessment of Shyness and Shyness Questionnaire

To assess the children's degree of shyness, we used the Inventory on Integrative Assessment of Child Temperament (German: *IKT—Inventar zur integrativen Erfassung des Kind-Temperaments*, Zentner, 2011). This is a standardized questionnaire that is widely used in clinical practice and is specifically designed for the age group addressed. The IKT has been validated with a normative sample of over 4,400 children, possesses convergent validity with equivalent English-language temperament diagnostics (e.g., the CBQ by Rothbart et al., 2001), and measures the temperament of 2- to 8-year-olds on five levels based on the integrative approach of Zentner and Bates (2008). With their approach, the authors pursue the goal of overcoming the manifold conceptions of child temperament research and providing a questionnaire that is valid across theories (Zentner and Bates, 2008). The levels comprise shyness (behavioral inhibition), susceptibility to frustration, activity level, attention span/task persistence and perceptual sensitivity. In our study, caregivers were asked during the first session (T1) to fill out the questionnaire and to estimate how often their child shows a described behavior using a 6-point Likert scale ranging from 1 (*never*) to 6 (*always*). This included behavioral aspects such as "hides behind her mother when she meets strangers." Based on the raw scores obtained from the responses to the questions, the evaluation procedure of the test requires a conversion into percentile ranks to allow an adequate interpretation of the child's temperament according to age and gender in relation to the normative sample of the test. The higher the percentile rank value, the shyer the child, with the minimum and maximum value being 0 and 100 respectively. In this vein, the IKT allows children

TABLE 1 | Mean participant characteristics for shy and nonshy children and standard deviation (SD).

Independent variable	Total (*N* = 28)	
	Nonshy (*n* = 18)	Shy (*n* = 10)
Age in years	*5.0 (0.5)*	*5.0 (0.4)*
Parental education level[1]	4.7 (*1.0*)	*4.1 (1.3)*
Gender		
Female	7 (39%)	4 (40%)
Male	11 (61%)	6 (60%)
SETK 3–5 sentence comprehension[2]	53.9 (*8.2*)	46.3 (*8.3*)
AWST-R expressive vocabulary	60.1 (*11.8*)	51.9 (*10.6*)
IKT shyness score[2]	40.4 (*24*)	87.2 (*6.3*)

[1]*Level of parental education on a scale from 1 (lowest) to 6 (highest).*
[2]*Converted raw values into percentile ranks.*

to be considered as notably shy if they have a clearly above-average score of over 75. Additionally, the shyness scale used in the test procedure provides a good internal consistency ($\alpha = 0.81$). In the normative sample, the agreement of both parents as a measure of interrater reliability was clearly above average ($r = 0.73$) and thus highest on the shyness level compared to the other levels (Zentner, 2011). Finally, in accordance with the evaluation procedure and based on the percentile ranks obtained, we grouped our sample into two levels: nonshy (n = 18) and shy (n = 10).

RESULTS

Table 1 presents an overview of all demographic data as well as the group means and standard deviations of the language and temperament measures.

Positive and Negative Shyness Reactions

A statistical analysis of shyness markers was not possible, because values tended toward zero in almost all training and testing periods. Therefore, we focused on analyzing the behavioral markers indicating pleasure and distress described in the following. In the *Limitations* section, we shall discuss some issues that may have led to the very rare occurrence of typical shyness reactions.

Expression of Pleasure

First, we wanted to know how far positive reactions were more likely in the nonshy group. We also assumed that due to the familiarity of the situation, positive reactions would increase over time in both groups, but especially in the shy group that would start off being more reserved but become more uninhibited over the course of the sessions.

Parametric statistical tests could not be used to analyze this dependent variable due to nonnormally distributed data and the small sample size. Therefore, we used the ANOVA type statistic (ATS)—a nonparametric equivalent of a mixed ANOVA (Akritas et al., 1997)—performed with the software R (package: nparLD, Noguchi et al., 2012). The ATS is regarded as a distribution-free test, but is mathematically more appropriate than classical rank-sum statistics such as Wilcoxon's rank-sum test. The test statistic is quite similar to ANOVA's F tests and exactly meets the α level while being conservative. It has been applied in developmental studies (Viertel, 2019; Tolksdorf et al., 2021). In addition, the ATS can tolerate unequal group sizes in the sample and is robust when studying longitudinal data because it considers their progression over time rather than comparisons between groups at each timepoint that may inflate type I error. The relative treatment effect (RTE) is a measure of the effect size and is estimated based on the actual sample. It can be determined for main effects as well as for interaction effects—that is, even for multifactorial designs with repeated measurements (Noguchi et al., 2012) as in the present study. The value of the relative effect RTE ranges between 0 and 1, whereby the occurrence of 0 and 1 means completely different conditions (e.g., for the shy and nonshy group); 0.5 indicates that the conditions do not differ at all (Brunner and Munzel, 2002; Noguchi et al., 2012).

A significant main effect demonstrated that nonshy children (Mdn = 1.14, IQR = 1.67, RTE = 0.57) used positive behaviors significantly more often than their shy peers (Mdn = 0.50, IQR = 1.37, RTE = 0.37), $F(1.00, 17.33) = 6.51$, $p < 0.05$, regardless of the situation (learning and testing). This supported Hypothesis H1a.

Additionally, the ATS revealed a highly significant main effect of time, $F(3.00, \infty) = 7.77$, $p < 0.001$. Positive reactions were highest at the beginning of the training of novel words (T1: Mdn = 1.71, IQR = 1.54, RTE = 0.61), decreased steadily over the course of training (T2: Mdn = 1.27, IQR = 1.06, RTE = 0.56; T3: Mdn = 0.63, IQR = 0.97, RTE = 0.40), and then remained relatively stable in both tests (Test 1: Mdn = 0.55, IQR = 1.15, RTE = 0.39; Test 2: Mdn = 0.60, IQR = 1.03, RTE = 0.40). Post hoc tests were applied with Bonferroni corrections. A significant decrease of positive reactions was detected from training T1 to Test 1 (difference = 34.0) and Test 2 (difference = 38.0) respectively, $\chi^2(4) = 15.86$, $p < 0.01$. In all cases, the critical difference was 33.21. This observed development of positive reactions contradicted Hypotheses H2a and H2b that positive reactions would increase with the repetition of a learning or test situation.

Moreover, contrary to our hypotheses, there was no interaction between time and shyness level, $F(3.00, \infty) = 0.65$, $p = 0.58$. Hence, shy children remained reserved over time by demonstrating fewer positive reactions such as smiles over the course of both the learning and the testing situations.

Expression of Distress

In a further step, we asked how far the frequency of negative reactions differed between shyness groups and training conditions depending on the time of training and testing situation. We hypothesized that the shy group would show negative reactions more frequently (H1b), but that these would decline more rapidly among the shy group than among the nonshy group in both the training situation (H2a) and the testing situation (H2b). Again, we refrained from using parametric statistics and used the ATS.

The ATS revealed a highly significant effect of time, $F(3.34, \infty) = 14.22$, $p < 0.001$. It was evident that negative reactions were stable across the first two training sessions (T1: Mdn = 0.45, IQR = 1.04, RTE = 0.58; T2: Mdn = 0.51, IQR = 0.75, RTE = 0.56), decreased during last training (T3: Mdn = 0.00, IQR = 0.32, *RTE* = 0.29), briefly increased during the first test (Test 1: *Mdn* = 0.57, *IQR* = 1.00, *RTE* = 0.62), and finally, flattened in the second test (Test 2: *Mdn* = 0.19, *IQR* = 0.20, *RTE* = 0.35).

There was also a trend toward a significant difference between shyness groups, with shy children (*Mdn* = 0.16, *IQR* = 0.50, *RTE* = 0.42) demonstrating negative reactions less frequently than nonshy children (*Mdn* = 0.33, *IQR* = 0.64, *RTE* = 0.55) in all training and test situations, $F(1.00, 18.21) = 3.33$, $p = 0.07$. Therefore, we rejected Hypothesis H1b.

However, both factors need to be interpreted in relation to each other, because there was a significant interaction effect between time and shyness group, $F(3.34, \infty) = 2.58$, $p < 0.05$. Post hoc tests showed that shy children expressed negative reactions less frequently, especially during the second training (*Mdn* = 0.24, *IQR* = 0.46, *RTE* = 0.41) as well as during the second test (*Mdn* = 0.12, *IQR* = 0.17, *RTE* = 0.26) compared to nonshy children (T2: *Mdn* = 0.66, *IQR* = 1.43, *RTE* = 0.71; Test 2: *Mdn* = 0.26, *IQR* = 0.19, *RTE* = 0.45). Differences during training ($W = 144.5$, $p < 0.01$, $r = -0.49$) and testing ($W = 146.5$, $p < 0.01$, $r = -0.51$) were both very significant. Thus, the result that only shy children showed negative reactions significantly less often when a situation was repeated supported Hypotheses H2a and H2b.

Additionally, in the shy group, multiple comparisons across time revealed that negative reactions decreased significantly from first to third training (T1 *vs.* T3: $\chi^2(4) = 16.17$, $p < 0.01$), with the observed difference of 20.5 exceeding the critical difference of 19.85. In the nonshy group, a significant reduction of negative reactions from T2 to T3 could be identified in the training situations, $\chi^2(4) = 19.34$, $p < 0.001$. The observed difference was 30.5 and thus higher than the critical difference of 26.63. Surprisingly, in the group of nonshy children, there was a significant increase in negative reactions from the last training session (T3) to the first testing session (Test 1) with an observed difference of 29.5. In the shy group, a tendency toward an increase was identified based on the observed difference of 19.5 (critical difference: 19.85). In conclusion, over the course of the training phase for novel words (T1–T3), negative reactions decreased in both groups, but this familiarization effect occurred more rapidly in the group of shy children.

Summarizing the frequency of both positive and negative reactions in the learning and test situations, it can be concluded that the group of shy children was generally less expressive compared to the group of less shy peers.

Shyness Score and Condition as Predictors of Word Learning

Finally, we focused on shy children's word learning, which we assumed would be less successful than that of nonshy children––but only during the first test session (H3). Children's word learning was measured based on the calculated linguistic composite score (0–3) reflecting their performance during the test tasks (cf. section *Assessment of Naming Performance*). As hypothesized, a strong trend could be observed during the first test, and shy children (*Mdn* = 0.51, *IQR* = 0.36, *range* = 0–1.52) tended to be less successful in retrieving the taught words than their nonshy peers (*Mdn* = 0.64, *IQR* = 0.27, *range* = 0–2.54), $W = 122$, $p = 0.06$, $r = -0.35$. Furthermore, also as hypothesized, both groups (shy: *Mdn* = 0.48, *IQR* = 0.27, *range* = 0–2.54; nonshy: *Mdn* = 0.72, *IQR* = 0.72, *range* = 0–2.06) did not differ significantly in their success at retrieving the words during the second test session ($W = 119.5$, $p = 0.16$, $r = -0.26$), which was a repetition of the first test a few days later.

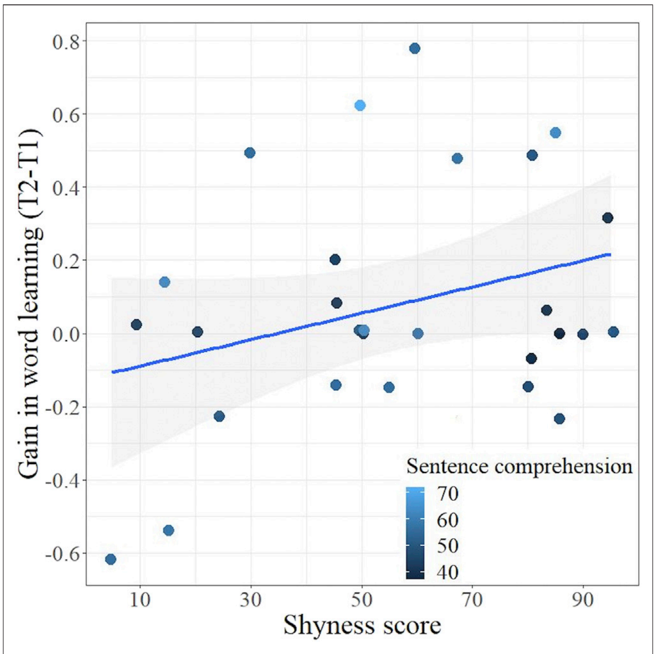


FIGURE 4 | Scatterplot with linear regression line (including 95% confidence interval) illustrating the predictive relation between level of shyness and gain in word learning (difference scores of word learning between T2 and T1). Receptive linguistic skills are integrated as converted percentile ranks.

Last of all, we wanted to go beyond the dichotomous group comparison and ask how the entire range of the shyness spectrum as an influencing factor predicted a gain in word learning when the children were already familiarized with the experienced testing situation during Test 1. As described, the testing situation took place in two different sessions (Test 1 and Test 2). Thus, we calculated the gain in word learning by measuring the difference scores between the first and the second test as a metric of learning gain. As a predictor variable, we did not use the categories *nonshy* and *shy*, but the calculated percentile ranks taken from the IKT described above (cf. section *Assessment of Shyness and Shyness Questionnaire*) that represented the full shyness spectrum and would allow us to take a closer look at the link between temperamental characteristics and word-learning processes.

Therefore, our hierarchical regression model included the shyness score as a predictor of a gain in word learning (first step). Moreover, when measuring children's word learning, it is known from the literature that existing linguistic knowledge should be taken into account because it contributes to word learning success (Stelmachowicz et al., 2004; McMurray et al., 2012). Therefore, in the second step, we integrated a language measure of children's receptive linguistic abilities (SETK 3–5 subtest sentence comprehension, cf. **Table 1**).

In Step 1, the model did not differ significantly from zero, $F(1, 26) = 2.72$, $p = 0.11$, with shyness level accounting for 9.48% of the variance in learning gain. In Step 2, the model approached statistical significance and accounted for an increased portion of gain in word learning, $F(2, 25) = 3.18$, $p = 0.06$, and could explain 20.28% of the variance. Shyness scores related significantly to a gain in word learning ($B = 0.005$, $t = 2.18$, $p = 0.04$) and uniquely contributed 14.63% to the total variance of our dependent variable (see **Figure 4**). The shyer the children were, the higher their growth of learning as suggested by the significant positive semipartial correlation ($r = 0.38$, $p = 0.05$). In terms of receptive linguistic abilities, we detected a nonsignificant trend ($B = 0.01$, $t = 1.84$, $p = 0.08$) toward a positive association between children's receptive language and word learning gain ($r = 0.33$, $p = 0.09$), whereas language abilities independently accounted for 10.96% of the variance. A comparison of both models confirmed that in Step 2, R^2 tended to change, $F(1, 25) = 3.38$, $p = 0.08$.

Finally, we checked whether other factors, which had not been considered in the multiple regression model before, were associated with word learning. Neither age ($r = 0.12$, $p = 0.27$, one-sided) nor expressive vocabulary collected by the AWST-R ($r = -0.02$, $p = 0.55$, one-sided) related significantly to a gain in word learning. In summary, this means that shy children showed greater gains in word learning than children who were less to moderately shy, but only when their receptive skills were considered in the model as well.

DISCUSSION

The present study examined how shy children, compared to nonshy children, enter into and maintain social interaction with a social robot during a word-learning educational setting, and how children's learning outcomes relate to their temperament. The study was motivated by previous research suggesting that shy children exhibit marked differences in their learning and social behavior toward interaction partners including social robots. In this respect, the contribution of our study is twofold: First, despite the importance of shy children's familiarization with a situation (Evans, 2001; Rimm-Kaufman and Kagan, 2005; Arbeau et al., 2010), we could not find any research investigating the behavior of shy children over a long-term interaction with a social robot. Therefore, we systematically assessed children's personality trait of shyness and investigated their behavior. More specifically, we specified markers of shyness and their signals of pleasure and distress toward the robot during multiple learning situations and during repeat testing situations in which children's learning was evaluated. Second, this study expands previous research by identifying different learning trajectories linked to the effects of shyness on children's learning performance. Whereas some work in human–human interaction indicates that shy children tend to perform more poorly in unfamiliar test procedures (Spere et al., 2004; Hilton and Westermann, 2017), findings on the impact of shyness on learning outcomes in the field of child–robot interaction are scarce.

Overall, results show that shy children not only interact differently with a robot compared to nonshy children, but also change their behavior over the course of the sessions. In fact, shy children interacted significantly less expressively with the robot in general. With regard to children's learning outcomes, shy children tended to score significantly lower on the first test, although they were able to close this gap during the second test, resulting in all children retrieving the learned words on a similar level. Surprisingly, we could even observe that once a certain familiarization with the test procedure was established, shyness related significantly to a gain in word learning when the receptive linguistic abilities were taken into account at the same time. In the following subsections, we shall interpret our findings one at a time.

Children's Expressions of Pleasure and Distress Toward the Robot

As expected, and in accordance with previous literature (Reddy, 2000; Putnam et al., 2006; Feinberg et al., 2012; Abe et al., 2014; Vogt et al., 2019; Tolksdorf et al., 2020b), we found that shy children were more reserved in their positive reactions toward the robot compared to their nonshy peers in all learning and testing situations. This lower level of positive reactions could be explained by the typical expressive pattern of shyness in novel social situations (Reddy, 2000; Colonnesi et al., 2014)—that is, shy children tend to display reduced emotional reactions and be more inhibited in unfamiliar social interactions (Poole and Schmidt, 2019). Surprisingly, we did not find support for our hypothesis that the number of positive reactions would increase with the repetition of a situation and increasing familiarity. Instead, our results revealed an opposite trend. Overall, the frequency of occurrences of positive reactions was most pronounced in the first two learning situations and then decreased steadily, reaching its lowest level in the final two test situations. With regard to the nonshy children, the results might be explained by the fact that precisely the novelty of the situation (e.g., the novel interaction partner and storytelling setting as suggested in Kanero et al., 2018) led to higher levels of engagement that were reflected in more positive reactions. In this vein, our results corroborate existing research in the area of child–robot interaction demonstrating that with the increasing duration of an interaction and repetitive behavior of a robot, children's engagement in terms of enjoyment could drop (de Wit et al., 2020). Thus, the repeated sessions with the robot might have led to a habituation effect and resulted in decreasing positive reactions. However, we could not observe a significant change in the expressiveness of the shy children's positive reactions over the course of the sessions, suggesting that their display of enjoyment of the situation remained constant on a low level, even during the learning situations.

Regarding the negative reactions as an expression of discomfort in a new situation, results show that behaviors such as frowning or narrowing the eyes reduced significantly more quickly over time in the group of shy children, probably due to the fact that they had become accustomed to the learning and testing situation. In particular, when the shy children were already familiarized with a specific setting (first learning situation or first test), reactions of distress were significantly lower in comparison to nonshy children. A decrease in reactions of distress also occurred in the nonshy children–but at a slower rate. In addition, in the latter group, the significant increase in negative reactions stood out when the setting (from learning to testing) and the associated demands on the children changed.

This finding has strong implications not only for studies evaluating social robots as interaction partners in educational settings, but also for studies examining word learning processes or other cognitive abilities in young children in general. Negative reactions (of distress) relate positively to general and social anxiety (Colonnesi et al., 2014), and negative shy reactions are associated with physiological changes that can activate the fight-or-flight system (Colonnesi et al., 2020) that consequently inhibit adaptive behavior and cognitive processes. During learning with others, they are often inhibited, which is reflected in deviant attentional processes (Hilton et al., 2019) as well as infrequent eye gaze to the other (Putnam et al., 2006). Thus, in our view, a warm-up that usually takes place a few minutes before the training or testing may not be sufficient for a specific population such as very shy children. Considering the fact that especially nonshy children expressed the highest proportion of reactions of distress during the first two learning sessions (T1 and T2), our conclusion is not exclusively of relevance for shy children, but also for children with different temperament traits.

Children's Word Learning

Our third hypothesis addressed how word learning in shy children differs from that in their nonshy peers at different time points. Motivated by the literature (Hilton and Westermann, 2017), we assumed that shy children would be less successful at retrieving the learned words than nonshy children, especially in the first test situation. We further hypothesized that the difference between groups would decline during the second test due to familiarization with the test procedure. At first impression, the performance of both shy and nonshy children in the word learning tests seems to be in line with our hypothesis, because in both tests, shy children were less successful in retrieving the trained words. Although the difference was marginally significant in the first test session, it disappeared in the second session. However, we considered that a further, more nuanced approach was needed, because there could be legitimate objections to our categorization into shy and nonshy. In particular, because children who were not at all shy fell into the same category as children who were on the threshold of being very shy (according to the questionnaire), we regarded our dichotomization as not being precise enough. Therefore, we conducted a regression analysis to determine the relation between the degree of shyness and word learning. Moreover, our multilevel model also took into account the children's linguistic knowledge, and we were able to show that, in addition to a shy temperament, the children's receptive abilities also tend to contribute to a success in word learning. Surprisingly, past studies on learning words with shy children (Hilton and Westermann, 2017; Hilton et al., 2019) or language learning studies with robots (Vogt et al., 2019) have not considered these linguistic abilities; therefore, our study marks another novelty in this field.

Interestingly, the children who were the shyest according to their parents made the largest gains in word learning, whereas those with the lowest shyness scores remained stable or even scored lower on the second test. This contrasts with results obtained in studies concluding that shy children are a) less likely to learn and retain new words (Hilton and Westermann, 2017) or b) have poorer productive vocabularies (e.g., Crozier and Hostettler, 2003). Additionally, these prior studies drew their findings from a single test—that is, word learning or vocabulary of shy children is typically assessed in a new environment with unknown people without any prior familiarization with the situation. This raises the question whether the shy children's test performance would be similar to that of nonshy children if they were already familiar with the situation (as realized in our study). As described above, shy children are afraid of being evaluated in unfamiliar situations, which consequently inhibits their performance, as evidenced by studies that a) examine vocabulary less invasively, for example by parents or in familiar school settings (Spere and Evans, 2009); or 2) in more anonymous group settings (Crozier and Hostettler, 2003); but also by studies that 3) determine the receptive vocabulary (Evans, 1996). These studies conclude that the linguistic performance of shy children does not differ from that of nonshy children. Therefore, based on our data, we assume that the shy children were more confident during the second testing and verbally expressed themselves more often once they were familiarized with the exact procedure and the demands of the test situation.

As well as the repetition of the testing situation, another possible reason could contribute to a short familiarization: the pragmatic frame of the situation of joint book reading that is structurally anchored in the test situation (Rohlfing et al., 2016). Accordingly, we can assume that the test situation recedes into the background or is not perceived as such by the children, so that their cognitive abilities can unfold (Rohlfing et al., 2016). Additionally, because our robot was introduced as a coequal peer (Kory Westlund et al., 2016; Vogt et al., 2017), it might not be perceived as an authoritative character, as other examiners are often perceived to be by shy children, but rather as an interaction partner who elicits learned words or, in general, verbal responses in a familiar situation. Therefore, shy children may feel less evaluated during an interaction with a robot, especially in terms of their performance, and this could lead them to be less cognitively inhibited and more confident when attempting to guess an answer.

LIMITATIONS

Finally, it is worth discussing why shyness markers appeared so rarely during training and testing that we were unable to carry out any analysis on this basis, but instead concentrated on behaviors expressing the children's emotionality of pleasure and distress but without any aversion. In 4-year-olds, state shyness is measured differently than in our study: Children are asked to sing a song on a stage in an unfamiliar location in front of a small group of strangers and their caregivers––the so-called performance task (Colonnesi et al., 2017). This scenario differs clearly from our setting in which the child is neither required to perform nor is her performance exposed to the other's attention. Hence, one objection might therefore be that the setting applied in our study failed to elicit shy reactions (Colonnesi et al., 2014). On the other hand, many factors triggering shyness reactions existed in our setting such as unknown interaction partners (and among them a robot), a new situation with novel words in an unfamiliar location, and the recall of a previously taught object of learning. Interestingly, the age of the children, which ranged from 48 to 70 months, may have contributed to the fact that (in particular negative) shyness reactions were rare–even in shy children. In this context, Colonnesi et al. (2020, p. 49) discuss their results with 72-month-olds: "A possible explanation is that later individual socio-cognitive development (e.g., advanced social cognition, social skills) and effortful control are responsible for less frequent and more regulated shy reactions." Instead, preschoolers express their shyness by reactions of distress and avoidance as well as ambivalence in their behavior–a pattern we were able to identify in our study as well. In this respect, Colonnesi et al. (2014) demonstrated that positive and negative reactions correlated positively with the corresponding shyness markers, thus concluding that they represent expressions of the same emotion but with a different emotional valence. In this context, it should be emphasized that we used a parental assessment of trait shyness based on a questionnaire. We regarded this as being the most reliable and valid for our study design rather than concentrating merely on state shyness measured in a specific situation. Finally, it is also important to recognize that one limitation of this study is its relatively small sample size. However, we wish to highlight that according to recent methodological findings, conducting long-term studies with small samples over multiple sessions while repeatedly measuring the variable of interest over time enhances replicability and robustness (Smith and Little, 2018). In this vein, the approach adopted here allows a particularly nuanced view of children's behavioral development. Additionally, the statistical procedures used were rather conservative in terms of determining significant effects and tolerant of both small sample sizes and unequal groups (Noguchi et al., 2012). Lastly, although the sample size used is consistent with previous studies utilizing a similar paradigm (McGregor et al., 2009) and we found clear differences between groups, more research is needed to further validate our findings.

CONCLUSIONS AND FUTURE CONSIDERATIONS

In conclusion, the results presented here offer new input for not only research with social robots in educational settings and the design of future learner–robot interactions but also for evaluating and measuring children's achieved learning outcomes. In fact, most past research in child–robot interaction has tested hypotheses by comparing average effects across the sample but ignored that effects may vary across individuals depending on existing intrinsic factors such as shyness. In this vein, the present study provides evidence that shy children do indeed demonstrate a distinctive behavior in terms of their interactions with a robot as well as their language learning. Our findings show that it is important to include the learner's temperamental characteristics, such as shyness, during child–robot interactions to inform the use of robots in the educational field. In this regard, current research strongly suggests the need to consider children's temperament in everyday practice in institutional settings such as kindergarten or school and provide a supportive climate for a variety of children and temperament types (Ann Sealy et al., 2021). Whereas approaches to automatically assessing a child's personality based on predefined behavioral traits have made substantial progress (Abe et al., 2017; Abe et al., 2020; Sano et al., 2020), the present study has taken first steps to determine which behaviors can actually be observed in shy children and how they develop in the long term over a period of several sessions.

From our results, we can derive some crucial aspects for designing future interactions with child-oriented social robots: On the one hand, testing shy children requires a greater familiarity with the situation, and this can be achieved by a prior acquaintance with the interaction partner, the location, and the items. For example, an extended warm-up session in advance of a learning *and* testing situation could be a solution in future interaction designs. This could be conducted individually in addition to carrying out an introduction in a group of children. Additionally, based on our data, we also suggest that future studies examining learning processes such as word learning should create scenarios in which children are tested at multiple time points, because this allows for a more precise focus on the specific learning processes, especially in shy children. On the other hand, given the high incidence of shyness as a normal variation in human personality in the overall population (Zimbardo et al., 1975; Kagan, 1994; Schmidt et al., 2020), our results also demonstrate that a nuanced assessment of shyness (e.g., parental assessment via a standardized questionnaire) is preferable, especially when it is central to the research question. In this vein, it is worth noting that temperamental characteristics are rarely collected in studies, and tests assessing children's linguistic and cognitive abilities are often administered by (almost) unfamiliar interaction partners.

We postulate that addressing children's individual differences and taking into account the personality of the interacting child can further guide future digital technologies and facilitate their integration into the educational landscape. However, at this point, we would like to be clear in our objective: We also see it as an ethical challenge to clarify whether future technologies should make automatic inferences about a child's temperament. Instead, it is important to gain further insights into how children's interactions with digital technologies such as social robots depend on their individual differences in order to enable educators and teachers to design future learning scenarios that allow all children to participate.

In future work, it would also be of interest to explore the effects of other dimensions of temperamental traits on interaction behavior such as susceptibility to frustration or attention span/

task persistence. This would shed further light on how individual adaptation to the learner and appropriate learning environments can be designed in the digital world of the future.

DATA AVAILABILITY STATEMENT

Based on the options within consent to future reuse of their data by other researcher that was given by the caregivers, the datasets presented in this article are not readily available and will not be made publicly available. Requests to access the datasets should be directed to Nils F. Tolksdorf, nils.tolksdorf@upb.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Review Board of Bielefeld University (EUB 2014-043). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

NT, FV, and KR conceived the study. NT and KR designed, and piloted the study. NT recruited participants. NT conducted the data collection. FV analyzed the data. NT, FV, and KR drafted the manuscript. All authors commented on, edited, and revised the manuscript prior to submission.

FUNDING

This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia and the research fund of the Faculty of Arts and Humanities of Paderborn University.

ACKNOWLEDGMENTS

We gratefully thank all the parents and children who participated in this study. We also thank our research assistants, Theresa Trieu and Dirk Hönemann, who were involved in the data annotation.

REFERENCES

Abe, K., Hieida, C., Attamimi, M., Nagai, T., Shimotomai, T., Omori, T., and Oka, N. (2014). "Toward playmate robots that can play with children considering personality," in *Proceedings of the second international conference on Human-agent interaction* (Tsukuba Japan: ACM), 165–168. doi:10.1145/2658861.2658913

Abe, K., Hamada, Y., Nagai, T., Shiomi, M., and Omori, T. (2017). "Estimation of Child Personality for Child-Robot Interaction," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).* (Lisbon: IEEE). 910–915. doi:10.1109/ROMAN.2017.8172411

Abe, K., Nagai, T., Hieida, C., Omori, T., and Shiomi, M. (2020). Estimating Children's Personalities through Their Interaction Activities with a Tele-Operated Robot. *J. Robot. Mechatron.* 32, 21–31. doi:10.20965/jrm.2020.p0021

Addison, T. L., and Schmidt, L. A. (1999). Are Women Who Are Shy Reluctant to Take Risks? Behavioral and Psychophysiological Correlates. *J. Res. Personal.* 33, 352–357. doi:10.1006/jrpe.1999.2249

Akritas, M. G., Arnold, S. F., and Brunner, E. (1997). Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs. *J. Am. Stat. Assoc.* 92, 258–265. doi:10.1080/01621459.1997.10473623

Ann Sealy, M., Moritz Rudasill, K., S. Barrett, J. J., Eum, J., Adams, N., Hinrichs, A., et al. (2021). "Temperament in the Early Elementary Classroom: Implications for Practice," in *Teacher Education In the 21st Century - Emerging Skills For a Changing World*. (London, United Kingdom: IntechOpen). doi:10.5772/intechopen.96270

Arbeau, K. A., Coplan, R. J., and Weeks, M. (2010). Shyness, Teacher-Child Relationships, and Socio-Emotional Adjustment in Grade 1. *Int. J. Behav. Dev.* 34, 259–269. doi:10.1177/0165025409350959

Asendorpf, J. B., and Meier, G. H. (1993). Personality Effects on Children's Speech in Everyday Life: Sociability-Mediated Exposure and Shyness-Mediated Reactivity to Social Situations. *J. Personal. Soc. Psychol.* 64, 1072–1083. doi:10.1037/0022-3514.64.6.1072

Barker, T. V., Reeb-Sutherland, B. C., and Fox, N. A. (2014). Individual Differences in Fear Potentiated Startle in Behaviorally Inhibited Children. *Dev. Psychobiol* 56, 133–141. doi:10.1002/dev.21096

Baxter, P., and Belpaeme, T. (2016). "A Cautionary Note on Personality (Extroversion) Assessments in Child-Robot Interaction Studies," in *2nd Workshop on Evaluating Child-Robot Interaction at HRI'16* (Christchurch, New Zealand: ACM).

Baxter, P., Ashurst, E., Read, R., Kennedy, J., and Belpaeme, T. (2017). Robot Education Peers in a Situated Primary School Study: Personalisation Promotes Child Learning. *PLOS ONE* 12, e0178126. doi:10.1371/journal.pone.0178126

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social Robots for Education: A Review. *Sci. Robot.* 3, eaat5954. doi:10.1126/scirobotics.aat5954

Boccanfuso, L., Scarborough, S., Abramson, R. K., Hall, A. V., Wright, H. H., and O'Kane, J. M. (2017). A Low-Cost Socially Assistive Robot and Robot-Assisted Intervention for Children with Autism Spectrum Disorder: Field Trials and Lessons Learned. *Auton. Robot* 41, 637–655. doi:10.1007/s10514-016-9554-4

Breazeal, C., Harris, P. L., DeSteno, D., Kory Westlund, J. M., Dickens, L., and Jeong, S. (2016). Young Children Treat Robots as Informants. *Top. Cogn. Sci.* 8, 481–491. doi:10.1111/tops.12192

Bruner, J. S. (1983). *Child's Talk: Learning to Use Language*. 1st ed. New York: W. W. Norton.

Brunner, E., and Munzel, U. (2002). Nichtparametrische Datenanalyse: Unverbundene Stichproben (Statistik und ihre Anwendungen). Berlin: Springer. doi:10.1007/978-3-642-56151-1

Calvo-Barajas, N., Perugia, G., and Castellano, G. (2020). "The effects of robot's facial expressions on children's first impressions of trustworthiness," in *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (Naples, Italy: IEEE). 165–171. doi:10.1109/RO-MAN47096.2020.9223456

Cao, H.-L., Esteban, P. G., Bartlett, M., Baxter, P., Belpaeme, T., Billing, E., et al. (2019). Robot-enhanced Therapy: Development and Validation of Supervised Autonomous Robotic System for Autism Spectrum Disorders Therapy. *IEEE Robot. Automat. Mag.* 26, 49–58. doi:10.1109/MRA.2019.2904121

Causo, A., Vo, G. T., Chen, I.-M., and Yeo, S. H. (2016). "Design of robots used as education companion and tutor," in Robotics And Mechatronics *Mechanisms and Machine Science*. Editors S. Zeghloul, M. A. Laribi, and J.-P. Gazeau (Cham: Springer International Publishing), 75–84. doi:10.1007/978-3-319-22368-1_8

Colonnesi, C., Bögels, S. M., de Vente, W., and Majdandžić, M. (2013). What Coy Smiles Say about Positive Shyness in Early Infancy. *Infancy* 18, 202–220. doi:10.1111/j.1532-7078.2012.00117.x

Colonnesi, C., Napoleone, E., and Bögels, S. M. (2014). Positive and Negative Expressions of Shyness in Toddlers: Are They Related to Anxiety in the Same Way? *J. Personal. Soc. Psychol.* 106, 624–637. doi:10.1037/a0035561

Colonnesi, C., Nikolić, M., de Vente, W., and Bögels, S. M. (2017). Social Anxiety Symptoms in Young Children: Investigating the Interplay of Theory of Mind and Expressions of Shyness. *J. Abnorm Child. Psychol.* 45, 997–1011. doi:10.1007/s10802-016-0206-0

Colonnesi, C., Nikolić, M., and Bögels, S. M. (2020). "Development and Psychophysiological Correlates of Positive Shyness from Infancy to Childhood," in *Adaptive Shyness: Multiple Perspectives On Behavior And Development*. Editors L. A. Schmidt and K. L. Poole (Cham: Springer International Publishing), 41–61. doi:10.1007/978-3-030-38877-5_3

Coplan, R. J., and Evans, M. A. (2009). At a Loss for Words? Introduction to the Special Issue on Shyness and Language in Childhood. *Inf. Child. Develop.* 18, 211–215. doi:10.1002/icd.620

Coplan, R. J., and Weeks, M. (2009). Shy and Soft-Spoken: Shyness, Pragmatic Language, and Socio-Emotional Adjustment in Early Childhood. *Inf. Child. Develop.* 18, 238–254. doi:10.1002/icd.622

Crozier, W. R., and Badawood, A. (2009). Shyness, Vocabulary and Children's Reticence in Saudi Arabian Preschools. *Inf. Child. Develop.* 18, 255–270. doi:10.1002/icd.623

Crozier, W. R., and Hostettler, K. (2003). The Influence of Shyness on Children's Test Performance. *Br. J. Educ. Psychol.* 73, 317–328. doi:10.1348/000709903322275858

de Wit, J., Brandse, A., Krahmer, E., and Vogt, P. (2020). "Varied human-like gestures for social robots: Investigating the effects on children's engagement and language learning" in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge United Kingdom: ACM), 359–367. doi:10.1145/3319502.3374815

Evans, M. A. (1996). Reticent Primary Grade Children and Their More Talkative Peers: Verbal, Nonverbal, and Self-Concept Characteristics. *J. Educ. Psychol.* 88, 739–749. doi:10.1037/0022-0663.88.4.739

Evans, M. A. (2001). "Shyness in the Classroom and Home," in *International Handbook of Social Anxiety: Concepts, Research and Interventions Relating to the Self and Shyness*. New York, US: John Wiley & Sons Ltd), 159–183.

Feinberg, M., Willer, R., and Keltner, D. (2012). Flustered and Faithful: Embarrassment as a Signal of Prosociality. *J. Personal. Soc. Psychol.* 102, 81–97. doi:10.1037/a0025403

Fischer, K., Jung, M., Jensen, L. C., and aus der Wieschen, M. V. (2019). "Emotion Expression in HRI - when and Why," in *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Daegu, Korea (South): IEEE), 29–38. doi:10.1109/HRI.2019.8673078

Grimm, H., Aktas, M., and Frevert, S. (2001). SETK 3 -5: Sprachentwicklungstest für drei- bis fünfjährige Kinder Göttingen: Hogrefe.

Grimminger, A., and Rohlfing, K. J. (2017). "Can You Teach Me?: Children Teaching New Words to a Robot in a Book Reading Scenario," in *Proc. WOCCI 2017: 6th International Workshop on Child Computer Interaction. (Glasgow: ISCA)*, 28–33. doi:10.21437/WOCCI.2017-5

Hilton, M., and Westermann, G. (2017). The Effect of Shyness on Children's Formation and Retention of Novel Word-Object Mappings. *J. Child. Lang.* 44, 1394–1412. doi:10.1017/S030500091600057X

Hilton, M., Twomey, K. E., and Westermann, G. (2019). Taking Their Eye off the Ball: How Shyness Affects Children's Attention during Word Learning. *J. Exp. Child Psychol.* 183, 134–145. doi:10.1016/j.jecp.2019.01.023

Horst, J. S. (2013). Context and Repetition in Word Learning. *Front. Psychol.* 4, 1–11. doi:10.3389/fpsyg.2013.00149

Ioannou, A., and Makridou, E. (2018). Exploring the Potentials of Educational Robotics in the Development of Computational Thinking: A Summary of Current Research and Practical Proposal for Future Work. *Educ. Inf. Technol.* 23, 2531–2544. doi:10.1007/s10639-018-9729-z

Kagan, J. (1994). *Galen's Prophecy: Temperament in Human Nature*. New York: Basic Books.

Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., and Göksun, T. (2018). Social Robots for Early Language Learning: Current Evidence and Future Directions. *Child. Dev. Perspect.* 12, 146–151. doi:10.1111/cdep.12277

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., and Belpaeme, T. (2017). "Child speech recognition in human-robot interaction," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna Austria: ACM)), 82–90. doi:10.1145/2909824.3020229

Kiese-Himmel, C. (2005). AWST-R-Aktiver Wortschatztest für 3- bis 5-jährige Kinder . Göttingen: Beltz Test.

Kory Westlund, J. M., Martinez, M., Archie, M., Das, M., and Breazeal, C. (2016). "Effects of Framing a Robot as a Social Agent or as a Machine on Children's Social Behavior," in 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), (New York, United States: IEEE), 688–693. doi:10.1109/ROMAN.2016.7745193

Levin, I. P., and Hart, S. S. (2003). Risk Preferences in Young Children: Early Evidence of Individual Differences in Reaction to Potential Gains and Losses. *J. Behav. Decis. Making* 16, 397–413. doi:10.1002/bdm.453

Manner, M. D., Gini, M., and Elison, J. (2018). "Graphically Representing Child-Robot Interaction Proxemics", in Workshop on Health Intelligence at AAAI Conference on Artificial Intelligence

Matsuda, Y.-T., Okanoya, K., and Myowa-Yamakoshi, M. (2013). Shyness in Early Infancy: Approach-Avoidance Conflicts in Temperament and Hypersensitivity to Eyes during Initial Gazes to Faces. *PLoS ONE* 8, e65476. doi:10.1371/journal.pone.0065476

McGregor, K. K., Rohlfing, K. J., Bean, A., and Marschner, E. (2009). Gesture as a Support for Word Learning: The Case of under. *J. Child. Lang.* 36, 807–828. doi:10.1017/S0305000908009173

McMurray, B., Horst, J. S., and Samuelson, L. K. (2012). Word Learning Emerges from the Interaction of Online Referent Selection and Slow Associative Learning. *Psychol. Rev.* 119, 831–877. doi:10.1037/a0029872

Nachtigäller, K., Rohlfing, K. J., and Mcgregor, K. K. (2013). A Story about a Word: Does Narrative Presentation Promote Learning of a Spatial Preposition in German Two-Year-Olds? *J. Child. Lang.* 40, 900–917. doi:10.1017/S0305000912000311

Noguchi, K., Gel, Y. R., Brunner, E., and Konietschke, F. (2012). nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments. *J. Stat. Softw.* 50, 1–23. doi:10.18637/jss.v050.i12

Oranç, C., and Küntay, A. C. (2020). Children's Perception of Social Robots as a Source of Information across Different Domains of Knowledge. *Cogn. Dev.* 54, 100875. doi:10.1016/j.cogdev.2020.100875

Park, H. W., Rosenberg-Kima, R., Rosenberg, M., Gordon, G., and Breazeal, C. (2017). "Growing growth mindset with a social robot peer," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. (Vienna Austria: ACM), 137–145. doi:10.1145/2909824.3020213

Poole, K. L., and Schmidt, L. A. (2019). Smiling through the Shyness: The Adaptive Function of Positive Affect in Shy Children. *Emotion* 19, 160–170. doi:10.1037/emo0000426

Putnam, S. P., Gartstein, M. A., and Rothbart, M. K. (2006). Measurement of Fine-Grained Aspects of Toddler Temperament: The Early Childhood Behavior Questionnaire. *Infant Behav. Dev.* 29, 386–401. doi:10.1016/j.infbeh.2006.01.004

Reddy, V. (2000). Coyness in Early Infancy. *Dev. Sci.* 3, 186–192. doi:10.1111/1467-7687.00112

Reddy, V. (2005). "Feeling Shy and Showing-Off: Self-Conscious Emotions Must Regulate Self-Awareness," in *Emotional Development: Recent Research Advances* (New York, NY, US: Oxford University Press), 183–204.

Rimm-Kaufman, S. E., and Kagan, J. (2005). Infant Predictors of Kindergarten Behavior: The Contribution of Inhibited and Uninhibited Temperament Types. *Behav. Disord.* 30, 331–347. doi:10.1177/019874290503000409

Rohlfing, K. J., and Grimminger, A. (2019). "Utilizing Pragmatic Frames as an analytical tool for children's performance during word learning," in *Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)* (Oslo, Norway: IEEE), 81–86. doi:10.1109/DEVLRN.2019.8850714

Rohlfing, K. J., Wrede, B., Vollmer, A.-L., and Oudeyer, P.-Y. (2016). An Alternative to Mapping a Word onto a Concept in Language Acquisition: Pragmatic Frames. *Front. Psychol.* 7, 1–18. doi:10.3389/fpsyg.2016.00470

Rohlfing, K. J., Leonardi, G., Nomikou, I., Raczaszek-Leonardi, J., and Hüllermeier, E. (2020). Multimodal Turn-Taking: Motivations, Methodological Challenges, and Novel Approaches. *IEEE Trans. Cogn. Dev. Syst.* 12, 260–271. doi:10.1109/TCDS.2019.2892991

Rohlfing, K. J., Grimminger, A., and Wrede, B. (2020). "The caregiver's role in keeping a child-robot interaction going," in *International perspectives on digital media and early literacy: the impact of digital devices on learning, language acquisition and social interaction*. Editors K. J. Rohlfing and C. Müller-Brauers (London: Routledge), 73–89. doi:10.4324/9780429321399-6

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., and Fisher, P. (2001). Investigations of Temperament at Three to Seven Years: The Children's Behavior Questionnaire. *Child. Dev.* 72, 1394–1408. doi:10.1111/1467-8624.00355

Rubin, K. H., Coplan, R. J., and Bowker, J. C. (2009). Social Withdrawal in Childhood. *Annu. Rev. Psychol.* 60, 141–171. doi:10.1146/annurev.psych.60.110707.163642

Salam, H., Celiktutan, O., Hupont, I., Gunes, H., and Chetouani, M. (2017). Fully Automatic Analysis of Engagement and its Relationship to Personality in Human-Robot Interactions. *IEEE Access* 5, 705–721. doi:10.1109/ACCESS.2016.2614525

Salter, T., Dautenhahn, K., and Boekhorst, R. t. (2006). Learning about Natural Human-Robot Interaction Styles. *Robotics Autonomous Syst.* 54, 127–134. doi:10.1016/j.robot.2005.09.022

Sano, T., Horii, T., Abe, K., and Nagai, T. (2020). "Explainable temperament estimation of toddlers by a childcare robot," in *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (Naples, Italy: IEEE). 159–164. doi:10.1109/RO-MAN47096.2020.9223574

Schmidt, L. A., and Buss, A. H. (2010). "Understanding Shyness: Four Questions and Four Decades of Research," in *The Development of Shyness and Social Withdrawal*. Editors K. H. Rubin and R. J. Coplan (New York: The Guilford Press), 23–41.

Schmidt, L. A., Poole, K. L., Fox, N. A., and Kagan, J. (2020). "The Study of Behavioral Inhibition and Temperamental Shyness across Four Academic Generations," in *"The Study of Behavioral Inhibition and Temperamental Shyness across Four Academic Generations," in* Adaptive Shyness. Editors L. A. Schmidt and K. L. Poole (Cham: Springer International Publishing), 3–21. doi:10.1007/978-3-030-38877-5_1

Schodde, T., Hoffmann, L., and Kopp, S. (2017). "How to manage affective state in child-robot tutoring interactions?," in *International Conference on Companion Technology (ICCT)* (Ulm, Germany: IEEE), 1–6. doi:10.1109/COMPANION.2017.8287073

Shiomi, M., Abe, K., Pei, Y., Ikeda, N., and Nagai, T. (2016). "I'm Scared : little children reject robots," in *Proceedings of the Fourth International Conference on Human Agent Interaction* (New York, NY: ACM), 245–247. doi:10.1145/2974804.2980493

Smith, P. L., and Little, D. R. (2018). Small Is Beautiful: In Defense of the Small-N Design. *Psychon. Bull. Rev.* 25, 2083–2101. doi:10.3758/s13423-018-1451-8

Smith Watts, A. K., Patel, D., Corley, R. P., Friedman, N. P., Hewitt, J. K., Robinson, J. L., et al. (2014). Testing Alternative Hypotheses Regarding the Association between Behavioral Inhibition and Language Development in Toddlerhood. *Child. Dev.* 85, 1569–1585. doi:10.1111/cdev.12219

Spere, K., and Evans, M. A. (2009). Shyness as a Continuous Dimension and Emergent Literacy in Young Children: Is There a Relation? *Inf. Child. Develop.* 18, 216–237. doi:10.1002/icd.621

Spere, K. A., Schmidt, L. A., Theall-Honey, L. A., and Martin-Chang, S. (2004). Expressive and Receptive Language Skills of Temperamentally Shy Preschoolers. *Inf. Child. Develop.* 13, 123–133. doi:10.1002/icd.345

Stelmachowicz, P. G., Pittman, A. L., Hoover, B. M., and Lewis, D. E. (2004). Novel-word learning in children with normal hearing and hearing Loss. *Ear & Hearing*, 25, 47–56. doi:10.1097/01.AUD.0000111258.98509.DE

Tolksdorf, N. F., and Mertens, U. (2020). ""Beyond words: children's multimodal responses during word learning with a robot," in *International perspectives on digital media and early literacy: the impact of digital devices on learning, language acquisition and social interaction*. Editors K. J. Rohlfing and C. Müller-Brauers (London: Routledge). 90–102. Available at: https://www.taylorfrancis.com/chapters/beyond-words-nils-tolksdorf-ulrich-mertens/e/10.4324/9780429321399-7.

Tolksdorf, N. F., and Rohlfing, K. J. (2020). "Parents' Views on Using Social Robots for Language Learning," in *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. (Naples, Italy: IEEE). 634–640. doi:10.1109/RO-MAN47096.2020.9223540

Tolksdorf, N. F., Siebert, S., Zorn, I., Horwath, I., and Rohlfing, K. J. (2020a). Ethical Considerations of Applying Robots in Kindergarten Settings: Towards an Approach from a Macroperspective. *Int. J. Soc. Robotics* 13, 129–140. doi:10.1007/s12369-020-00622-3

Tolksdorf, N. F., Viertel, F., and Rohlfing, K. J. (2020b). "Do shy children behave differently than non-shy children in a long-term child-robot interaction?," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (Cambridge United Kingdom: ACM)*, 488–490. doi:10.1145/3371382.3378367

Tolksdorf, N. F., Crawshaw, C. E., and Rohlfing, K. J. (2021). Comparing the Effects of a Different Social Partner (Social Robot vs. Human) on Children's Social Referencing in Interaction. *Front. Educ.* 5, 569615. doi:10.3389/feduc.2020.569615

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., and Leseman, P. (2019). Social Robots for Language Learning: A Review. *Rev. Educ. Res.* 89, 259–295. doi:10.3102/0034654318821286

Vaughn, S., and La Greca, A. (1992). "Beyond Greetings and Making Friends: Social Skills from a Broader Perspective," in *Contemporary Intervention Research In Learning Disabilities: An International Perspective*. Editor B. Y. L. Wong (New York, NY: Springer New York), 96–114. doi:10.1007/978-1-4612-2786-1_6

Viertel, F. E. (2019). Training of visual perspective-taking (level 1) by means of role reversal imitation in 18-20-montholds with particular regard to the temperamental trait shyness. Paderborn, Germany: Paderborn University Library. doi:10.17619/UNIPB/1-889

Vogt, P., de Haas, M., de Jong, C., Baxter, P., and Krahmer, E. (2017). Child-robot Interactions for Second Language Tutoring to Preschool Children. *Front. Hum. Neurosci.* 11, 1–6. doi:10.3389/fnhum.2017.00073

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., et al. (2019). "Second language tutoring using social robots: a large-scale study," in *2019 14th ACM/IEEE International Conference On Human-Robot Interaction (HRI)* (Daegu, Korea (South): IEEE), 497–505. doi:10.1109/HRI.2019.8673077

Vollmer, A.-L., Read, R., Trippas, D., and Belpaeme, T. (2018). Children Conform, Adults Resist: A Robot Group Induced Peer Pressure on Normative Social Conformity. *Sci. Robot.* 3, eaat7111. doi:10.1126/scirobotics.aat7111

Walters, M. L., Dautenhahn, K., te Boekhorst, R., Kheng Lee Koay, Kheng., Kaouri, C., Woods, S., et al. (2005). "The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment," in *IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*, 2005. Nashville (TN, USA: IEEE), 347–352. doi:10.1109/ROMAN.2005.1513803

Zentner, M., and Bates, J. E. (2008). Child Temperament: An Integrative Review of Concepts, Research Programs, and Measures. *Int. J. Dev. Sci.* 2, 7–37. doi:10.3233/DEV-2008-21203

Zentner, M. (2011). *Inventar zur integrativen Erfassung des Kind-Temperaments*. Bern: Huber Verlag.

Zimbardo, P. G., Pilkonis, P. A., and Norwood, R. (1975). Social Disease Called Shyness. *Psychol. Today* 8, 69

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 in Robotics and AI

ORIGINAL RESEARCH
published: 30 June 2021
doi: 10.3389/frobt.2021.693050

Machine Teaching for Human Inverse Reinforcement Learning

Michael S. Lee, Henny Admoni and Reid Simmons*

Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, United States

As robots continue to acquire useful skills, their ability to teach their expertise will provide humans the two-fold benefit of learning from robots and collaborating fluently with them. For example, robot tutors could teach handwriting to individual students and delivery robots could convey their navigation conventions to better coordinate with nearby human workers. Because humans naturally communicate their behaviors through selective demonstrations, and comprehend others' through reasoning that resembles inverse reinforcement learning (IRL), we propose a method of teaching humans based on demonstrations that are informative for IRL. But unlike prior work that optimizes solely for IRL, this paper incorporates various human teaching strategies (e.g. scaffolding, simplicity, pattern discovery, and testing) to better accommodate human learners. We assess our method with user studies and find that our measure of test difficulty corresponds well with human performance and confidence, and also find that favoring simplicity and pattern discovery increases human performance on difficult tests. However, we did not find a strong effect for our method of scaffolding, revealing shortcomings that indicate clear directions for future work.

Keywords: inverse reinforcement learning, learning from demonstration, scaffolding, policy summarization, machine teaching

Edited by:
Tony Belpaeme,
Ghent University, Belgium

Reviewed by:
Goren Gordon,
Tel Aviv University, Israel
Emmanuel Senft,
University of Wisconsin-Madison, United States

****Correspondence:***
Michael S. Lee
ml5@andrew.cmu.edu

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: *09 April 2021*
Accepted: *07 June 2021*
Published: *30 June 2021*

Citation:
Lee MS, Admoni H and Simmons R (2021) Machine Teaching for Human Inverse Reinforcement Learning. Front. Robot. AI 8:693050. doi: 10.3389/frobt.2021.693050

1 INTRODUCTION

As robots become capable in tasks once accomplished only by humans, the extent of their influence will depend in part on their ability to teach and convey their skills. From the youngest of us learning to handwrite (Sandygulova et al., 2020; Guneysu Ozgur et al., 2020) to practitioners of crafts such as chess, many of us stand to benefit from robots that can effectively teach their mastered skill. Furthermore, our ability to collaborate fluently with robots partly depends our understanding of their behaviors. For example, workers at a construction site could better coordinate with a new delivery robot if the robot could clearly convey its navigation conventions (e.g. when it would choose to go through mud over taking a long detour).

While demonstrations are a natural method of teaching and learning behaviors for humans, its effectiveness still hinges on conveying an informative set of demonstrations. The literature on how humans generate and understand behaviors provides insight into what makes a demonstration informative. Cognitive science suggests that humans often model one another's behavior as exactly or approximately maximizing a reward function (Jern et al., 2017; Jara-Ettinger et al., 2016; Lucas et al., 2014), which they can infer through reasoning resembling inverse reinforcement learning (IRL) (Ng and Russell, 2000; Jara-Ettinger, 2019; Baker et al., 2009; Baker et al., 2011). Furthermore, humans are often able to obtain a behavior that (approximately) maximizes a reward function through planning, which can be modeled as dynamic programming or Monte Carlo tree search (Shteingart and Loewenstein, 2014; Wunderlich et al., 2012).

Putting these insights together, we can often expect humans to be able to model others' behaviors once equipped with their reward functions.[1] For example, upon seeing a new worker consistently arrive on time to each workday, a manager will infer that the worker places high values on punctuality and consistency and will arrive promptly at other work-related functions. Thus, the problem of conveying a behavior or skill can be reduced to conveying the underlying reward function, and the informativeness of a demonstration can be quantified by how much information it reveals regarding the reward function using IRL.

Though IRL offers a principled measure of a demonstration's informativeness, human learning is multi-faceted and is also influenced by other factors, such as the simplicity of explanations (Lombrozo, 2016). Thus, unlike prior work on machine teaching that optimizes solely for IRL (Brown and Niekum, 2019), this paper incorporates insights on how humans effectively learn to further accommodate human learners.

In this work, we explore whether augmenting IRL with insights from human teaching improves human learning over optimizing for IRL alone. We first employ *scaffolding* from social constructivism (learning theory) to encourage demonstrations that are not just informative but also comprehensible. Specifically, we assume a general human learner without prior knowledge, and sequence demonstrations that incrementally increase in informativeness and difficulty. Noting the cognitive science literature that suggests humans favor simple explanations that follow a discernible pattern (Lombrozo, 2016; Williams et al., 2010), we also optimize for visual *simplicity and pattern discovery* when selecting demonstrations. Finally, toward effective *testing* of the learner's understanding, we show that the measure of a demonstration's informativeness during teaching can be inverted into a measure of expected difficulty for a human to predict that exact demonstration during testing.

Two user studies strongly correlate our measure of test difficulty with human performance and confidence, with low, medium, and high difficulty tests yielding high, medium, and low performance and confidence respectively. Study results also show that favoring simplicity and pattern discovery significantly increases human performance on difficult tests. However, we do not find a strong effect for our method of scaffolding, revealing shortcomings that indicate clear directions for future work.

2 RELATED WORK

2.1 Policy Summarization and Machine Teaching

The problem of policy summarization considers which states and actions should be conveyed to help a user obtain a global understanding of a robot's policy (i.e. behavior or skill) (Amir et al., 2019). There are two primary approaches to this problem. The first relies on heuristics to evaluate the value of communicating certain states and actions, such as entropy (Huang et al., 2018), differences in Q-values (Amir and Amir, 2018), and differences between the policies of two agents (Amitai and Amir, 2021).

We build on the second approach, which follows the machine teaching paradigm (Zhu et al., 2018). Given an assumed learning model of the student (e.g. IRL to learn a reward function), the machine teaching objective is to select the minimal set of teaching examples (i.e. demonstrations) that will help the learner arrive at a specific target model (e.g. a policy). Though machine teaching was first applied to classification and regression (Zhu, 2015; Liu and Zhu, 2016), it has also recently been employed to convey reward functions from which the corresponding policy can be reconstructed. Huang et al. (2019) selected informative demonstrations for humans modeled to employ approximate Bayesian IRL for recovering the reward. This technique requires the true reward function to be within a candidate set of reward functions over which to perform Bayesian inference, and computation scales linearly with the size of the set. Cakmak and Lopes (2012) instead focused on IRL learners and selected demonstrations that maximally reduced uncertainty over all viable reward parameters, posed as a volume removal problem. Brown and Niekum (2019) improved this method (particularly for high dimensions) by solving an equivalent set cover problem instead with their Set Cover Optimal Teaching (SCOT) algorithm. However, SCOT is not explicitly designed for human learners and this paper builds on SCOT to address that gap.

2.2 Techniques for Human Teaching

Human teaching and learning is a multifaceted process that has been studied extensively. Thus, we also take inspiration from social constructivism (learning theory) and cognitive science in informing how a robot may teach a skill to a human learner so that the learner may correctly reproduce that skill in new situations.

Scaffolding: Scaffolding is a well-established pedagogical technique in which a more knowledgeable teacher assists a learner in accomplishing a task currently beyond the learner's abilities, e.g. by reducing the degrees of freedom of the problem and/or by demonstrating partial solutions to the task (Wood et al., 1976). Noting the benefits seen by automated scaffolding to date [e.g. Sampayo-Vargas et al. (2013)], we implement the first recommendation made by Reiser (2004) for software-based scaffolding, which is to reduce the complexity of the learning problem through additional structure. Specifically, we incorporate this technique when teaching a skill by providing demonstrations that sequentially increase in informativeness and difficulty.

Simplicity and Pattern Discovery: Studies on explanations preferred by humans indicate a bias toward those that are simpler and have fewer causes (Lombrozo, 2016). Furthermore, Williams et al. (2010) found that explanations can be detrimental if they do not help the learner to notice useful patterns or even mislead them with false patterns. Together, these two works support the idea that explanations should minimize distractions that

[1] Ng and Russell (2000) suggest that "the reward function, rather than the policy, is the most succinct, robust, and transferable definition of the task."

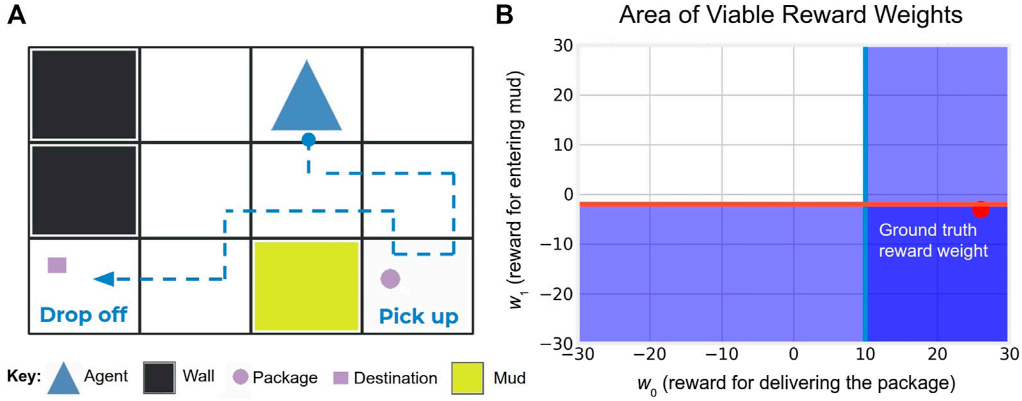


FIGURE 1 | (A) A demonstration $\mathcal{D}$ of an optimal policy π in the delivery domain. Agent aims to deliver the package to the destination while avoiding walls and avoiding mud if the detour is not too costly. **(B)** The left demonstration can be translated into a set of half-space constraints on the underlying policy reward weights using **Eq. 4**. The darker shaded region is where all constraints (the red and light blue lines) hold true, which corresponds to the behavior equivalence class BEC($\mathcal{D}|\pi$), see Section 3.3.

potentially inspire false correlations and instead highlight and reinforce the minimal set of causes. We thus also optimize for simplicity and pattern discovery when selecting demonstrations that naturally "explain" the underlying skill.

Testing: Effective scaffolding requires an accurate diagnosis of the learner's current abilities to provide the appropriate level of assistance throughout the teaching process (Collins et al., 1988). A common diagnostic method is presenting the learner with tests of varying difficulties and assessing their understanding of a skill. Toward this, we propose a way to quantify the difficulty of a test that specifically assesses the student's ability to predict the right behavior in a new situation.

3 TECHNICAL BACKGROUND

3.1 Markov Decision Process

The robot's environment is represented as an instance (indexed by i) of a deterministic[2] Markov decision process, $MDP_i := (\mathcal{S}_i, \mathcal{A}, T_i, R, \gamma, S_i^0)$, where $\mathcal{S}_i$ and $\mathcal{A}$ denote the state and action sets, $T_i : \mathcal{S}_i \times \mathcal{A} \rightarrow \mathcal{S}_i$ the transition function, $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ the reward function, $\gamma \in [0, 1]$ the discount factor, and S_i^0 the initial state distribution, and $\mathcal{S} : \cup_i \mathcal{S}_i$ the union over the states of all related instances of MDPs, which we call a domain (to be described in the following paragraphs).

Finally, the robot has an optimal policy (i.e. a skill) $\pi_i^\star : \mathcal{S}_i \rightarrow \mathcal{A}$ that maps each state in an MDP to the action that will optimize the reward in an infinite horizon. A sequence of (s_i, a, s_i') tuples obtain by following $\pi^\star$ gives rise to an optimal trajectory (i.e. a demonstration) $\xi^\star$, where $s_i, s_i' \in \mathcal{S}_i, a \in \mathcal{A}$. We assume that R can be expressed as a weighted linear combination of l reward features[3] $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^l$, i.e. $R = \mathrm{w}^{\star\top} \phi(s, a, s')$ (Abbeel and Ng, 2004). We also assume that the human is aware of all aspects of an MDP (including the reward features) but not the weights $\mathrm{w}^\star$.

Let a domain refer to a collection of related MDPs that share $\mathcal{A}, R, \gamma$ but differ in $\mathcal{S}_i$, T_i and S_i^0. Take for example the delivery domain, which modifies the Taxi domain (Dietterich, 1998) by adding mud (see **Figure 1**). The robot is rewarded for efficiently delivering the package to the destination while avoiding the mud if the detour is not too costly. Though MDPs in this domain may vary in the number and locations of mud patches and subsequently offer a diverse set of demonstrations (e.g. see **Figure 2**), they importantly share the same reward function R.

Because instances of a domain share R, the various demonstrations all support inference over the same $\mathrm{w}^\star$ through IRL. Thus, we overload the notation $\pi^\star$ to refer to any policy of a domain instance that optimizes a reward with $\mathrm{w}^\star$. Furthermore, while a demonstration strictly consists of both an optimal trajectory $\xi^\star$ (obtained by following $\pi^\star$) and the corresponding MDP (minus $\mathrm{w}^\star$), we will refer to a demonstration only by $\xi^\star$ in this work for notational simplicity.

Having represented the robot's environment and policy, we now define the problem of generating demonstrations for teaching that policy through the lens of machine teaching.

3.2 Machine Teaching for Policies

As formalized by Lage et al. (2019), machine teaching for policies seeks to convey a set of demonstrations $\mathcal{D}$ of size n (i.e. the allotted budget for teaching set) that will maximize the similarity ρ between $\pi^\star$ and the policy $\hat{\pi}$ recovered using a model $\mathcal{M}$ on $\mathcal{D}$

$$\arg\max_{\mathcal{D}\in\Xi} \rho(\hat{\pi}(\mathcal{D}, \mathcal{M}), \pi^\star) \text{ s.t. } |\mathcal{D}| = \text{n} \quad (1)$$

where Ξ is the set of all optimal demonstrations of $\pi^\star$ in a domain. We assume that the $\mathcal{M}$ employed by humans to approximate the underlying $\mathrm{w}^\star$ is IRL. Once $\mathrm{w}^\star$ (and the subsequent reward function) is approximated, we assume that human learners are able to arrive at $\pi^\star$, i.e. the skill, through planning on the underlying MDP.

[2]Though we assume a deterministic MDP, the methods described here naturally generalize to MDPs with stochastic transition functions and policies.

[3]This assumption can be made without loss of generality as the reward features can be nonlinear with respect to states and actions and be arbitrarily complex.

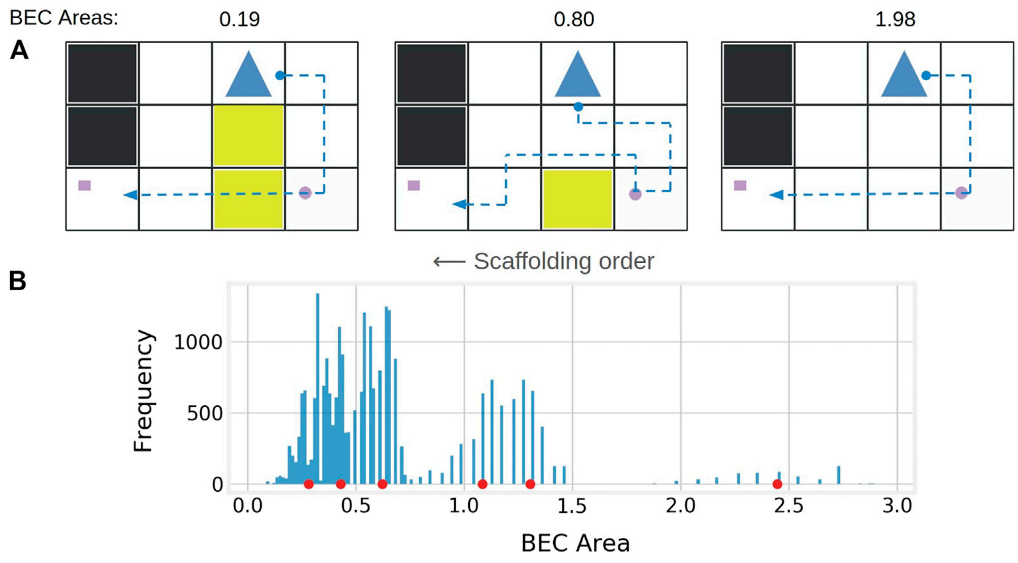


FIGURE 2 | (A) Sample demonstrations exhibiting scaffolding, simplicity, and pattern discovery. We scaffold by showing demonstrations that incrementally decrease in BEC area (which appears to correlate inversely with informativeness and difficulty). Simplicity is encouraged by minimizing visual clutter (i.e. unnecessary mud patches). Pattern discovery is encouraged by holding the agent and passenger locations constant while highlighting the single additional toll between demonstrations that changes the optimal behavior. **(B)** Histogram of BEC areas of the 25,600 possible demonstrations in the delivery domain. Cluster centers returned by k-means (k = 6) are shown as red circles along the *x*-axis. Demonstrations from every other cluster are selected and shown in order of largest to smallest BEC area for scaffolded machine teaching.

Thus, the teaching objective reduces to effectively conveying w^* through well-selected demonstrations.[4] In order to quantify the information a demonstration provides on w^*, we leverage the idea of behavior equivalence classes.

3.3 Behavior Equivalence Class

The *behavior equivalence class* (BEC) of π is the set of (viable) reward weights under which π is still optimal. The larger the BEC(π) is, the greater the potential uncertainty over w^* that is underlying the robot's optimal policy.

$$\mathrm{BEC}(\pi) = \left\{ w \in \mathbb{R}^l \mid \pi \text{ optimal w.r.t. } R = w^\top \phi(s, a, s') \right\} \quad (2)$$

The BEC(π) can be calculated as the intersection of the following half-space constraints generated by the central IRL equation (Ng and Russell, 2000)

$$\begin{gathered} w^\top \left(\mu_\pi^{(s,a)} - \mu_\pi^{(s,b)} \right) \geq 0 \\ \forall a \in \underset{a' \in \mathcal{A}}{\arg\max}\, Q^*(s, a'), b \in \mathcal{A}, s \in \mathcal{S} \end{gathered} \quad (3)$$

where $\mu_\pi^{(s,a)} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \phi(s_t) \middle| \pi, s_0 = s, a_0 = a \right]$ is the vector of expected reward feature counts accrued from taking action a in s, then following π after, and $Q^*(s, a)$ refers to the optimal Q-value in a state and a possible action (Watkins and Dayan, 1992).

Brown and Niekum (2019) proved that the BEC($\mathcal{D}|\pi$) of a set of demonstrations $\mathcal{D}$ of a policy π can be formulated similarly as the intersection of the following half-spaces

$$w^\top \left(\mu_\pi^{(s,a)} - \mu_\pi^{(s,b)} \right) \geq 0, \ \forall (s, a) \in \mathcal{D}, b \in \mathcal{A}. \quad (4)$$

Using the **Eq. 4**, every demonstration can be translated into a set of constraints on the viable reward weights.

Consider an example in the delivery domain with $A = \{$*up*, *down*, *left*, *right*, *pick up*, *drop*, *exit*$\}$, $w^* = [26, -3, -1]$[5] and binary reward features $\phi = [$*dropped off package at destination*, *entered mud*, *action taken*$]$. The demonstration in the left image of **Figure 1** corresponds to the constraints in the right image. With a unit cost for each action, the constraints on viable reward weights intuitively indicate that 1) $w_0^* \geq 10$ since a total of 10 actions were taken in the demonstration and that 2) $w_1^* \leq -2$ as the detour around the mud took two actions.

3.4 Set Cover Optimal Teaching (SCOT)

SCOT (Brown and Niekum, 2019) allows a robot to select the minimum number of demonstrations that results in the smallest BEC area (i.e. the intersection of the constraints) for an IRL learner. As it only considers IRL, it serves as a baseline method to the techniques proposed in this work that augment SCOT with human teaching strategies.

[4]In principle, a robot could simply convey w^* explicitly to a human. However, it can be nontrivial for humans to map precise numerical reward weights to the corresponding optimal behavior through planning, especially if there is large number of reward features. Thus, providing demonstrations that inherently carry information regarding w^* and directly conveying the optimal behavior can be more a effective teaching method for human learners.

[5]In practice, we also require that $\|w\|_1 = 1$ to circumvent the scaling invariance of IRL solutions and to eliminate the degenerate all-zero reward function (Brown and Niekum, 2018). We convey the non-normalized w here for intuition.

The SCOT algorithm is summarized here for completeness. The robot first translates all possible demonstrations of its policy in a domain into a corresponding set of BEC constraints. After taking a union of these constraints, redundant constraints are removed using linear programming (Paulraj and Sumathi, 2010). These non-redundant constraints together form the minimal representation of BEC(π^*). SCOT now iteratively runs through all possible demonstrations again and greedily adds to the teaching set $\mathcal{D}$ the demonstration that covers as many of the remaining constraints in BEC(π^*), until all constraints are covered.[6] These steps correspond to lines 2–13 in **Algorithm 1**.

4 PROPOSED TECHNIQUES FOR TEACHING HUMANS

4.1 Scaffolding

The SCOT algorithm efficiently selects the minimum number of demonstrations that results in the smallest BEC area for a pure IRL learner (Brown and Niekum, 2019). Such a learner is assumed to fully grasp these few highly nuanced examples that delicately straddle decision-making boundaries and find any other demonstrations redundant. However, *we posit that the BEC area of a demonstration not only inversely corresponds to the amount of information it contains about the possible values of* w^**, but also inversely corresponds to the effort required for a human to extract that information.* Thus humans will likely benefit from additional scaffolded examples that ease them in and incrementally relax the degrees of freedom of the learning problem.

We develop a scaffolding method for a learner without any prior knowledge, outlined as follows. First, obtain the SCOT demonstrations that contains the maximum information on w^*. If space remains in the teaching budget n for additional demonstrations, begin scaffolding by sorting all possible demonstrations in a domain according to their BEC areas. Then cluster them using k-means into twice as many clusters as the remaining budget to ensure that no two consecutive demonstrations are nearly identical in BEC area (see **Figure 2**). Randomly draw m candidate demonstrations from every other cluster. Finally from these n pools of candidate demonstrations, select the ones that best optimize visuals for the teaching set $\mathcal{D}$ (as described in the next section). See lines 16–21 in **Algorithm 1**. In this paper, the algorithm always divided the BEC areas into 6 clusters, considering every other cluster to correspond to “low”, “medium”, and “high” information respectively.

4.2 Simplicity and Pattern Discovery

Though the BEC area of a demonstration provides an unbiased, quantitative measure of the information transferred to a pure IRL learner, *human learners are likely also influenced by the medium of the demonstration, e.g. visuals, and the simplicity and patterns it affords*. For example, visible differences between sequential demonstrations can highlight relevant aspects, while visual clutter that does not actually influence the robot’s behavior (e.g. extraneous mud not in the path of the delivery robot) may distract or even mislead the human.

We perform a greedy sequential optimization for pattern discovery and then for simplicity. We first encourage pattern matching by considering candidates from different BEC clusters (which often exhibit qualitatively different behaviors) that are most visually similar to the previous demonstration.[7] The aim is to highlight a change in environment (e.g. a new mud patch) that caused the change in behavior (e.g. robot takes a detour) while keeping all other elements constant. We then optimize for simplicity. A measure of visual simplicity is manually defined for each domain (e.g. the number of mud patches in the delivery domain), and out of the scaffolding candidates, the visually simplest demonstration is selected.

The proposed methods for scaffolding and visual optimization come together in **Algorithm 1**.[8] Since the highest information SCOT demonstrations are selected first then demonstrations are selected via k-means clustering from high to low information, the algorithm concludes by reversing the demonstration list to order the demonstrations from easiest to hardest (line 28).[9] $\widehat{\mathrm{N}}[\cdot]$ denotes the operation of extracting unit normal vectors corresponding to a set of half-space constraints, and \ denotes set subtraction. An example of a sequence of demonstrations that exhibits scaffolding, simplicity, and pattern discovery can be found at the top of **Figure 2**.

4.3 Testing

An optimal trajectory’s BEC area intuitively captures its informativeness as a teaching demonstration. The smaller the area, the less uncertainty there is regarding the value of w^*.

We propose a complementary and novel idea: *that the BEC area can be inverted as a measure of a trajectory’s difficulty as a question during testing*, i.e. when a human is asked to predict the robot’s trajectory in a new situation. Intuitively, a large BEC area indicates that there are many viable reward weights for a demonstration, and thus the human does not need to precisely understand w^* to correctly predict the robot’s trajectory. We can also use this measure to scaffold tests of varying difficulties to gauge the human’s understanding of w^* and subsequently π^*.

[6]Instead of greedily adding the first demonstration that covers the most remaining constraints of BEC(π^*) at each iteration, one can enumerate all possible combinations of demonstrations that cover BEC(π^*) and optimize for simplicity and pattern discovery here as well.

[7]We measure the visual similarity of two states by defining a hash function over a domain’s state space and calculating the edit distance between the two corresponding state hashes.

[8]An implementation is available at https://github.com/SUCCESS-MURI/machine-teaching-human-IRL.

[9]In theory, one could order SCOT and k-means demonstrations jointly by BEC area and potentially allowing them to mix in order. However, a SCOT demonstration that contributes a maximally informative constraint of BEC(π^*) may in fact have a large BEC area. Thus, showing this SCOT demonstration early on may actually render a later k-means demonstration as uninformative (i.e. the SCOT demonstration’s BEC(π^*) constraint may cause a later k-means demonstration’s constraints to be redundant). Instead, showing k-means demonstrations that iteratively decrease in BEC area, then showing SCOT demonstrations ensures that the learner receives non-redundant constraints on w^* at each step.

Algorithm 1 Machine Teaching for Human Learners.

Require: π^*: optimal policy, $\mathbb{D}$: set of all MDPs belonging to a domain, Ξ: all possible demonstrations of π^* in a domain, n: teaching budget, m: cluster pool size
// Obtain SCOT demos
$U = \varnothing$
for $MDP \in \mathbb{D}$ **do**
  $U = U \cup \hat{\mathrm{N}}[\mathrm{BEC}(\pi^*)]$ ▷ Obtain BEC(π^*) using Eq. 3 on each MDP comprising a domain
end for
$U = \mathrm{removeRedundantConstraints}(U)$ ▷ See (Paulraj and Sumathi, 2010)
$\mathcal{D} = [\,], C = \varnothing$
while $|U \setminus C| \neq 0$ **do**
  $\xi^* = \arg\max_{\xi \in \Xi} |\hat{\mathrm{N}}[\mathrm{BEC}(\xi|\pi^*)] \cap (U \setminus C)|$ ▷ Eq. 4
  $\mathcal{D}.\mathrm{append}(\xi^*)$
  $C = C \cup \hat{\mathrm{N}}[\mathrm{BEC}(\xi|\pi^*)]$
  $\Xi = \Xi \setminus \xi^*$
end while
// Select candidates to fill teaching budget via scaffolding
if $|\mathcal{D}| < n$ **then**
  $\mathbf{D}_{cand} = \varnothing$ ▷ Set of sets
  $\Xi_{sorted} = \mathrm{sortByIncreasingBECArea}(\Xi)$
  $\Xi_{cluster} = \mathrm{kMeans}\big(\Xi_{sorted}, 2(n - |\mathcal{D}|)\big)$
  for $\big(i = 1, i = 2(n - |\mathcal{D}|), i \mathrel{+}= 2\big)$ **do**
    $\mathbf{D}_{cand} = \mathbf{D}_{cand} \cup \{\mathrm{sampleTraj}(m, \Xi_{cluster}[i])\}$
  end for
  // Downselect from candidates based on visuals
  for $\mathcal{D}_{cand} \in \mathbf{D}_{cand}$ **do**
    $\mathcal{D}_{prelim} = \mathrm{maximizeVisualSimilarity}(\mathcal{D}_{cand}, \mathcal{D})$
    $\xi^* = \mathrm{maximizeVisualSimplicity}(\mathcal{D}_{prelim})$
    $\mathcal{D}.\mathrm{append}(\xi^*)$
  end for
  $\mathcal{D} = \mathrm{reverse}(\mathcal{D})$ ▷ Order demonstrations from easiest to hardest
end if
return $\mathcal{D}$ ▷ Final demonstration set to show human

5 USER STUDIES

We ran two online user studies that involved participants watching demonstrations of a 2D agent's policy and predicting the optimal trajectory in new test environments.[10] The studies were designed to evaluate the following hypotheses.

H1: The BEC area of a demonstration correlates 1) inversely to the expected difficulty for a human to correctly predict it during testing, and 2) directly to their confidence in that prediction.

H2: The BEC area of a demonstration also correlates 1) inversely to the information transferred to a human during teaching and 2) inversely to the subsequent test performance.

H3: Forward scaffolding (demonstrations shown in increasing difficulty) will result in better qualitative assessments of the teaching set and better participant test performance over no scaffolding (only high difficulty demonstrations shown) and backward scaffolding (demonstrations shown in decreasing difficulty), in that order.

H4: Positive visual optimization will result in better qualitative assessments of the teaching set and better test performance over negative visual optimization (with positive and negative visual optimization corresponding to the maximization and minimization, respectively, of both simplicity and pattern discovery).

The two user studies jointly tested H1. The first study tested H2 and the second study tested H3 and H4.

5.1 Domains

Three simple gridworld domains were designed for this study (see **Figure 3**). The available actions were {*up*, *down*, *left*, *right*, *pick up*, *drop*, *exit*}. Each domain consisted of one shared reward feature of unit action cost, and two unique reward features as follows.

Delivery domain: The agent is rewarded for bringing a package to the destination and penalized for moving into mud.

Two-goal domain: The agent is rewarded for reaching one of two goals, with each goal having a different reward.

Skateboard domain: The agent is rewarded for reaching the goal. It is penalized less per action if it has picked up a skateboard (i.e. riding a skateboard is less costly than walking).

To convey an upper bound on the positive reward weight, the agent exited from the game immediately if it encountered an environment where working toward the positive reward would yield a lower overall reward (e.g. too much mud along its path). The semantics of each domain were masked with basic geometric shapes and colors to prevent biasing human learners with priors.

[10]Code for the user studies, videos of teaching and testing demonstrations, and the collected data are available at https://github.com/SUCCESS-MURI/psiturk-machine-teaching.

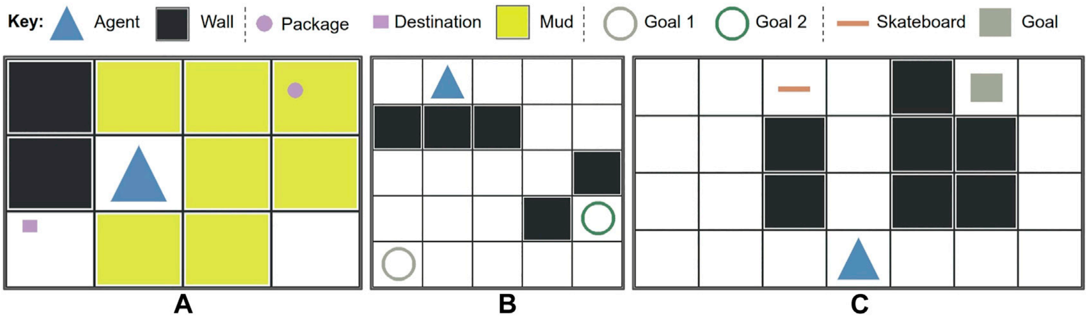


FIGURE 3 | Three domains were presented in the user study, each with a different set of reward weights to infer from demonstrations using inverse reinforcement learning. **(A)** delivery, **(B)** two-goal, **C**: skateboard.

All domains were implemented using the simple_rl framework (Abel, 2019).

5.2 Study Design

The first and second user studies (US1 and US2, respectively) used the same domains, procedures, and measures, though they differed in which variable was manipulated.

US1 explored how BEC area of demonstrations correlates with a human's understanding of the underlying policy. Thus, the between-subjects variable was *information class*, with three levels: low, medium, and maximum (i.e. SCOT). The low and medium information demonstrations were selected from the fifth and third BEC clusters respectively (see **Figure 2**). When selecting multiple demonstrations from a *single* cluster, we optimized for visual simplicity and *dissimilarity* as diversity[11] of demonstrations has been shown to improve human learning (Amir and Amir, 2018; Huang et al., 2019). The number of demonstrations shown in each domain was set to equal the number of SCOT demonstrations for fair comparison (2 for delivery and skateboard, 3 for two-goal).

US2 explored how incorporating human learning strategies impacts a human's understanding of the underlying policy. Specifically, it examined how the presence and direction of scaffolding, and optimization of visuals, would impact the human's test performance. The between-subjects variables were *scaffolding class* (none, forward, and backward), and *visual optimization* (positive and negative). For scaffolding class, forward scaffolding showed demonstrations according to **Algorithm 1**, backward scaffolding showed forward scaffolding's demonstrations in reverse, and no scaffolding showed all high informative examples from the 1st BEC cluster (**Figure 2**). Five demonstrations were shown for each domain, always ending with demonstrations determined by SCOT.

Both US1 and US2 had two additional within-subject variables: *domain* (delivery, two-goal, and skateboard, described in **Section 5.1**) and *test difficulty* (low, medium, and high, determined by the BEC area of the test).

[11]Note that **Algorithm 1** already achieves diversity by scaffolding demonstrations across *different* BEC clusters and thus benefits instead from visual similarity.

For both user studies, participants first completed a series of tutorials that introduced them to the mechanics of the domains they would encounter. In the tutorials, participants learned that the agent would be rewarded or penalized according to key events (i.e. reward features) specific to each domain. They were then asked to generate a few predetermined trajectories in a practice domain with a live reward counter to familiarize themselves with the keyboard controls and a practice reward function. Finally, participants entered the main user study and completed a single trial in each of the delivery, two-goal, and skateboard domains. Each trial involved a teaching portion and a test portion. In the teaching portion, participants watched videos of optimal trajectories that maximized reward in that domain, then answered subjective questions about the demonstrations (M2-M4, see **Section 5.3**). In the subsequent test portion, participants were given six new test environments and asked to provide the optimal trajectory. The tests always included two low, two medium, and two high difficulty environments shown in random order. For each of the tests, participants also provided their confidence in their response (M5). The teaching videos for each condition were pulled from a filtered pool of 3 exemplary sets of demonstrations proposed by **Algorithm 1** to control for bias in the results. The tests were likewise pulled from a filtered pool of 3 exemplary sets of demonstrations for each of the low, medium, and high difficulty test conditions.

Finally, though the methods described in this paper are designed for a human with no prior knowledge regarding any of the weights, the agent in our user studies assumed that the human was aware of the step cost and only needed to learn the relationship between the remaining two weights in each domain. This simplified the problem at the expense of a less accurate human model and measure of a demonstration's informativeness via BEC area. However, the effect was likely mitigated in part by the clustering and sampling in **Algorithm 1**, which only makes use of coarse BEC areas.

5.3 Measures

The following objective and subjective measures were recorded to evaluate the aforementioned hypotheses.

M1. Optimal response: For each test, whether the participant's trajectory received the optimal reward or not was recorded.

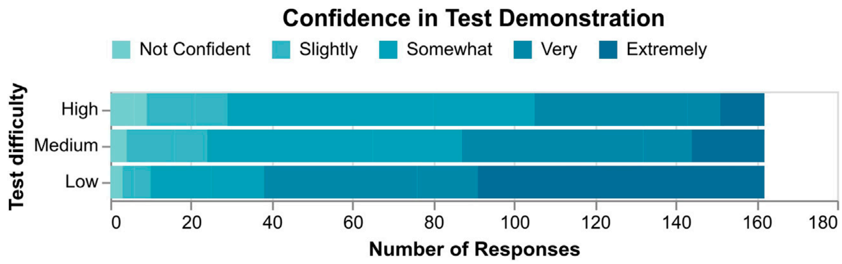


FIGURE 4 | Participants were significantly more confident of their responses as test difficulty decreased.

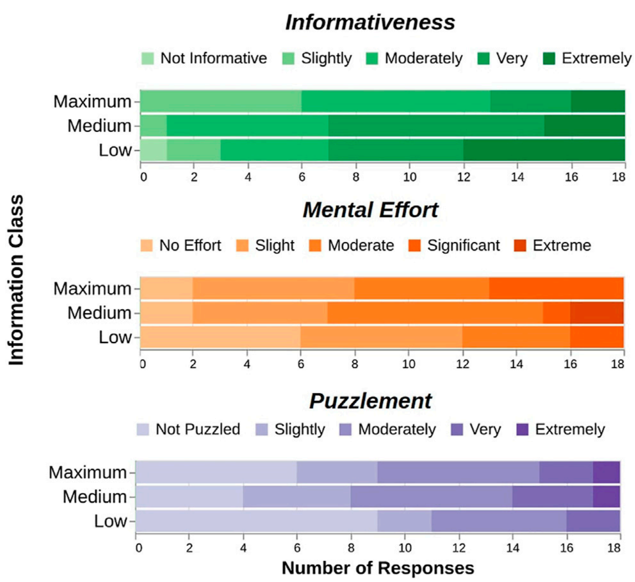


FIGURE 5 | The information class of demonstrations only significantly influences their perceived informativeness, ironically decreasing from low to maximum information class. This suggests that a demonstration's intrinsic information content (as measured by its BEC area) does not always correlate with the information transferred to human learners. No significant effects were found between information class and mental effort or puzzlement.

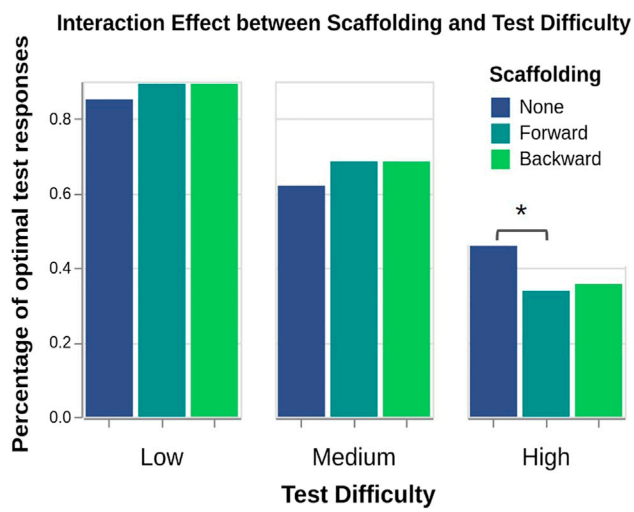


FIGURE 6 | Though the three scaffolding conditions perform similarly in aggregate across all tests, "no scaffolding" significantly increases performance for high difficulty tests.

M2. Informativeness rating: 5-point Likert scale with prompt "How informative were these demonstrations in understanding how to score well in this game?"

M3. Mental effort rating: 5-point Likert scale with prompt "How much mental effort was required to process these demonstrations?"

M4. Puzzlement rating: 5-point Likert scale with prompt "How puzzled were you by these demonstrations?"

M5. Confidence rating: 5-point Likert scale with prompt "How confident are you that you obtained the optimal score?"

6 RESULTS

One hundred and sixty two participants were recruited using Prolific (Palan and Schitter, 2018) for the two user studies. Participants' ages ranged from 18 to 57 (M = 26.07, SD = 8.35). Participants self-reported gender (roughly 67% male, 30% female, 2% non-binary, and 1% preferred to not disclose). Each of the nine possible between-subjects conditions across the two user studies were randomly assigned 18 participants (such that US1 and US2 contained 54 and 108 participants respectively), and the order of the domains presented to each participant was counterbalanced.

The three domains were designed to vary in the difficulty of their respective optimal trajectories. We calculated an intraclass coefficient (ICC) based on a mean-rating (k = 3), consistency-based, 2-way mixed effects model (Koo and Li, 2016) to evaluate the consistency of each participant's performance across domains. A low ICC value of 0.37 ($p < .001$) indicated that performance in fact varied considerably across domains for each participant. We subsequently average each participant's scores across the domains in all following analyses, potentially yielding results that are representative of domains with a range of difficulties.

H1: We combine the test responses from both user studies as they shared the same pool of tests. A one-way repeated measures ANOVA revealed a statistically significant difference

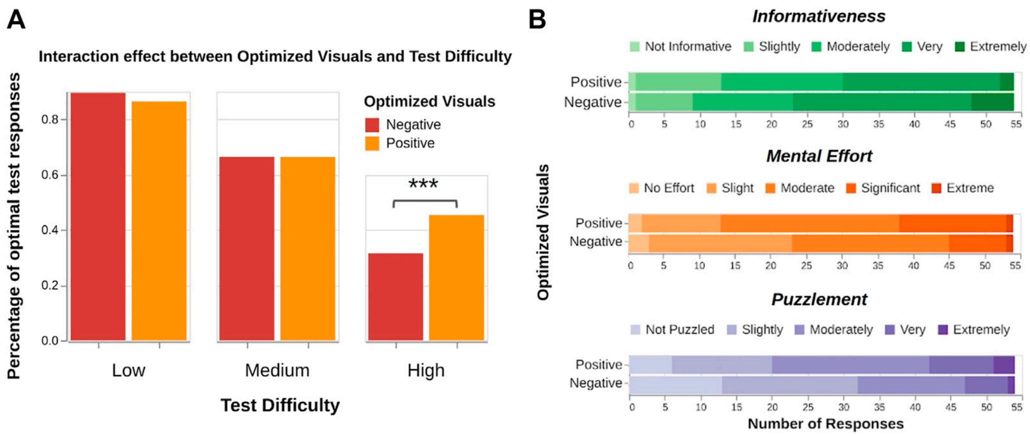


FIGURE 7 | (A) Optimizing teaching demonstration visuals does not significant affect performance on low and medium difficulty tests, but leads to a significant improvement on high difficulty tests. **(B)** Ratings on mental effort and puzzlement surprisingly increased for positive visual optimization, likely an artifact of unforeseen study design effects. No significant effects were found for ratings on informativeness.

in the percentage of optimal responses (M1) across test difficulty $(F(2, 322) = 275.35, p < .001)$. Post-hoc pairwise Tukey analyses further revealed significant differences between each of the three groups, with the percentage of optimal responses dropping from low $(M = 0.89)$, to medium $(M = 0.68)$, to high $(M = 0.36)$ test difficulties $(p < .001$ in all cases).

Spearman's rank-order correlation further showed a significant inverse correlation between test difficulty and confidence (M5, $r_s = -.40, p < .001, N = 486$). See **Figure 4** for the raw confidence data.

Objective and subjective results both support H1, that BEC area can indeed be used as a measure of difficulty for testing. We thus proceed with the rest of the analyses with "test difficulty" as a validated independent variable.

H2: A two-way mixed ANOVA on percentage of optimal responses (M1) did not reveal a significant effect of information class of the teaching set $(F(2, 51) = 1.23, p = .30)$, though test difficulty had a significant effect consistent with the H1 analysis $(F(2, 102) = 118.58, p < .001)$. There was no interaction between information class and test difficulty $(F(4, 102) = 0.67, p = .61)$.

Spearman's correlation test only found a significant negative correlation between information class and perceived informativeness (M2, $r_s = -0.28, p = .04, N = 54$). Neither mental effort (M3, $p = .08$) nor puzzlement (M4, $p = .36$) were found to have significant correlations with information class. See **Figure 5** for the raw subjective ratings.

The data failed to support H2. The data suggests that IRL alone is indeed an imperfect model of human learning, motivating the use of human teaching techniques to better accommodate human learners.

There was no correlation between information class and test performance, likely a result of two factors. First, the number of demonstrations provided (two or three) across the conditions in US1 were likely too few for human learners, who are not pure IRL learners and can sometimes benefit from "redundant" examples that reinforce a concept. Second, as will be discussed under the scaffolding subsection in **Section 7.2**, BEC area is likely an insufficient model of a demonstration's informativeness to a human and warrants further iteration.

Accordingly, maximum information demonstrations provided by SCOT $(M = 0.61)$ failed to significantly improve the percentage of optimal responses compared to medium $(M = 0.65)$ and low $(M = 0.67)$ information demonstrations as IRL would have predicted. The subjective results further indicate that people ironically found the maximally informative demonstrations least informative. We hypothesize that participants struggled to digest the information contained within SCOT's demonstrations all at once, motivating the use of scaffolding to stage learning into mangeable segments.

H3: A two-way mixed ANOVA on percentage of optimal responses (M1) revealed a significant interaction effect between scaffolding and test difficulty $(F(4, 210) = 2.79, p = .03)$. Tukey analyses showed that no scaffolding $(M = 0.46)$ yielded significantly better test performance than forward scaffolding $(M = 0.34)$ for high difficulty tests $(p = .05)$. Though not statistically significant, a trend of forward and backward scaffolding outperforming no scaffolding on low $(M = 0.89, 0.89, 0.85$ respectively) and medium difficulty tests $(M = 0.69, 0.69, 0.62$ respectively) can be observed as well (see **Figure 6**).

A two-way mixed ANOVA surprisingly did not reveal a significant effect from scaffolding $(F(2, 105) = 0.02, p = .98)$ but did find a significant effect for test difficulty $(F(2, 210) = 167.63, p < .001)$ on percentage of optimal responses (M1) as expected.

A Kruskal–Wallis test did not find differences between the informativeness $(H(2) = 5.18, p = .07)$, mental effort $(H(2) = 1.16, p = .56)$, or puzzlement $(H(2) = 0.59, p = .74)$ ratings (M2–M4) of differently scaffolded teaching sets.

The data largely failed to support H3. Forward and backward scaffolding surprisingly led to nearly identical test performance. Though no scaffolding performed similarly overall, it yielded a significant increase in performance specifically for high difficulty

TABLE 1 | Coding of qualitative participant responses as resembling inverse reinforcement learning (IRL) or imitation learning (IL), or "unclear."

Learning style	Raw counts (across user studies)		Percentages (across coders)	
	Coder 1	Coder 2	User study 1 (%)	User study 2 (%)
IRL	25	27	32	68
IL	7	9	27	12
Unclear	15	11	41	20

tests. These two surprising results are addressed in the discussion. The subjective measures did not indicate any clear relationships.

H4: A two-way mixed ANOVA on percentage of optimal responses (M1) revealed significant effects of test difficulty ($F(2, 212) = 169.21, p < .001$) and an interaction effect between optimized visuals and test difficulty ($F(2, 212) = 5.61, p = .004$). Exploring the interaction effect with Tukey analyses revealed that visual optimization had no effect on test performance on low ($p = .24$) and medium ($p = .90$) difficulty tests, but led to a significant improvement in performance in high ($p < .001$) difficulty tests for positive visual optimization ($M = 0.45$) over negative ($M = 0.31$), see **Figure 7**. The two-way mixed ANOVA did not reveal a significant from optimized visuals alone ($F(1, 106) = 2.27, p = .13$).

A Mann–Whitney U test surprisingly found that ratings for mental effort ($U(N_{neg} = 54, N_{pos} = 54) = 1131.5, p = .03$) and puzzlement ($U(N_{neg} = 54, N_{pos} = 54) = 1082.5, p = .02$) (M3 and M4) increased for positive visual optimization. Informativeness ratings were not found to differ significantly between the two visual optimizations ($p = .11$).

The data partially supports H4. Optimizing visuals improved test performance for high difficulty tests. However, optimizing visuals also yielded counterintuitive results for the subjective measures on mental effort and puzzlement, which we address in the following section.

7 DISCUSSION

7.1 Learning Styles

Analyzing the free-form comments provided by participants throughout the user studies revealed unexpected insights about their learning styles. Though this paper assumed that participant learning would only resemble IRL, we discovered it sometimes resembled imitation learning[12], which models humans as learning the optimal behavior directly from demonstrations (as opposed to through an intermediate reward function like IRL) (Daw et al., 2005; Lage et al., 2019). For example, one participant expounded upon their mental effort Likert rating (M3) with following description of IRL-style learning: "You need to make a moderate amount of mental effort to understand all the rules and outweight [sic] everything and see what is worth it or not in the game." In contrast, another expounded upon their used mental effort rating with the following description of IL-style learning: "The primary 'mental effort' was in memorizing the patterns of each level/stage and matching the optimal movements for them."

To better understand the types of learning employed by our participants, we analyzed their optional responses to the following questions: "Feel free to explain any of your selections above if you wish:" (asked in conjunction with prompts for ratings of informativeness, mental effort, and puzzlement of demonstrations in each domain, i.e. up to three times) and "Do you have any comments or feedback on the study?" (asked after the completion of the full study, i.e. once). Similar to Lage et al. (2019), we coded relevant responses from participants regarding their thought process as resembling IRL (e.g. "So, the yellow squares should be avoided if possible and they possibly remove two points when crossed but I'm not sure") or as resembling IL (e.g. "I did not understand the rule regarding yellow tiles. It seems they should be avoided, but not always. Interesting. . ."), or as "unclear" (e.g. "After some examples I feel like I'm understanding way better these puzzles."). A second coder uninvolved in the study independently labeled the same set of responses, assigning the same label to 79% of the responses. A Cohen's kappa of 0.64 between the two sets of codings further indicates moderate to substantial agreement (Landis and Koch, 1977; Altman, 1990; McHugh, 2012). Please refer to the **Supplementary Material** for the responses, labels, and further details on the coding process.

As **Table 1** conveys, both coders agreed that more responses resembled IRL than IL and "unclear" combined, suggesting that people perhaps employed IRL more often than not. However, we note that the way the tutorials introduced the domains may have influenced this result. For example, explicitly conveying each domain's unique reward features and clarifying that a trajectory's reward is determined by a weighting over those features may have encouraged participants to first infer the reward weights from optimal demonstrations (e.g. through IRL) and then infer the optimal policy (as opposed to directly inferring the optimal policy e.g. through IL).

Examining the percentage of each response across the two user studies reveals another interesting trend. Responses were far more likely to be coded as IRL in US2, where participants got to see five demonstrations as opposed to US1, where participants only got to see two or three demonstrations. This echoes the observation of Lage et al. (2019) that people may be more inclined to use IL over IRL in less familiar situations, which may be moderated in future studies

[12]Note that the term "behavior cloning" is sometimes used instead to refer to the process of directly learning the optimal behavior. Accordingly, "imitation learning" is sometimes used to refer to the broad class of techniques that learn optimal behavior from demonstrations, encompassing both behavior cloning and IRL (Osa et al., 2018).

through more extensive pre-study practice and/or additional informative demonstrations that better familiarize the participant to the domains.

Finally, out of 15 participants who provided more than one response, coders agreed that eight appeared to employ the same learning style throughout the user study (e.g. participants 129 and 142 in US2 only provided responses resembling IRL), four appeared to have changed styles through the user study (e.g. participants 59 in US1 and 20 in US2 provided various responses that resembled IL, IRL, or were unclear), and three were ambiguous (i.e. one coder coded a consistent learning style while the other did not). Though we controlled for learning effects by counterbalancing the order of the domains, participants likely found the domains to vary in the difficulty of their respective optimal trajectories (as suggested by the ICC score). Furthermore, certain conditions led to significant differences in subjective and objective outcomes (e.g. maximum information demonstrations were ironically perceived to be least informative (H2) and positive visual optimization improved performance for high difficulty tests (H4)). We thus hypothesize that the varying difficulties in domains and conditions non-trivially influenced the learning styles at different times [e.g. by moderating familiarity (Lage et al., 2019)].

Future work: The multi-faceted nature of human learning can be described by a number of models such as IRL and IL. Lage et al. (2019) show post hoc that tailoring the teaching to the human's favored learning style can improve the learning outcome. Thus, predicting a human's current learning style a priori or in situ (e.g. by using features such as the human's familiarity of the task or domain) and matching the teaching appropriately in real time will be an important direction of future work.

7.2 Scaffolding

Though BEC area is a well-motivated preliminary model of a demonstration's informativeness to a human, backward scaffolding's unexpected on-par performance with forward scaffolding suggests that it is insufficient and our scaffolding order likely was not clear cut in either direction. In considering possible explanations, we note that **Eq. 4** presents a computationally elegant method of generating BEC constraints via sub-optimal, one-step deviations from the optimal trajectory. However, these suboptimal trajectories do not always correspond to the suboptimal trajectories in the human's mind (e.g. which may allow more than one-step deviations). This sometimes leads to a disconnect between a demonstration's informativeness as measured by BEC area and its informativeness from the point of view of the human.

Furthermore, forward and backward scaffolding (each comprised of low, medium, and high information demonstrations) yielded higher performance for low and medium difficulty tests, and no scaffolding (comprised of only high information demonstrations) yields significantly higher performance for high difficulty tests. Improved performance when matching the informativeness and difficulty of teaching and testing demonstrations respectively (which yields similar demonstrations) further suggests that IL-style learning may have also been at play.

Finally, participants across each condition never achieved a mean score of greater than 0.5 for high difficulty tests, indicating that they were largely unable to grasp the more subtle aspects of the agent's optimal behavior. While the five demonstrations shown in US2 should have conveyed the maximum possible information (in an IRL-sense), they were not as effective in reality. One reason may be that human cognition is constrained by limited time and computation (Griffiths, 2020), and at times may opt for approximate, rather than exact, inference (Vul et al., 2014; Huang et al., 2019). Approximate inference (and even IL-style learning) indeed would have struggled with high difficulty tests whose optimal behavior could often only be discerned through exact computation of rewards. In addition to potentially showing more demonstrations (including "redundant" demonstrations that reinforce concepts and are still useful for approximate IRL), we believe that more effective scaffolding that further simplifies the concepts being taught while simultaneously challenging human's current knowledge will be key to addressing this gap, as we discuss next.

Future work: We propose two directions for future work on scaffolding. First, we note that our selected demonstrations often revealed information about multiple reward weights at once, which could be difficult to process. Instead, we can further scaffold by teaching about one weight at a time, when possible. Second, Reiser (2004) suggests that scaffolding should not only provide structure that reduces problem complexity but at times induce cognitive conflict to challenge and engage the learner. The current method of scaffolded teaching assumes that the learner has no prior knowledge when calculating a demonstration's informativeness (e.g. **Algorithm 1** considers a repeat showing of a demonstration to a learner to be equally as informative as the first showing). But when filtering for teaching and testing sets for the user studies, we sometimes observed and accounted for the fact that demonstrations with the same BEC area could further vary in informativeness or difficulty to different learners based on whether it presented an expected behavior or not. We believe that providing demonstrations which incrementally deviate from the human's current model will be more informative to a human and would be better suited to scaffolding.

7.3 Simplicity and Pattern Discovery

Optimizing visuals improved test performance, but only for high difficulty tests. This suggests that simplicity and pattern discovery could produce a meaningful reduction in complexity for only high information demonstrations (which contain the insights necessary to do well on the high difficulty tests), while those of low and medium information were already comprehensible.

We found counterintuitive results on mental effort or puzzlement ratings (M3–M4) for H4, where ratings for mental effort and puzzlement increased from negative to positive visual optimizations. One factor may have been the open-ended phrasing of the corresponding Likert prompts that failed to always elicit the intended measure. For example, one participant expounded upon their mental effort rating by saying "it takes a bit of efford [sic] remembering that you can quit at any time," referencing the difficulty of remembering all available actions rather than the intended difficulty of performing inference over the optimal behavior.

Similarly, the open-ended prompt for puzzlement failed to always query specifically for potential puzzlement arising from (a potentially counterintuitive) ordering of the demonstrations. Instead it sometimes invited comments such as 'I think i [sic] saw the same distance to the objective 2 times and 2 differnt [sic] outcomes,' and interestingly informed us of possible unforeseen confounders on puzzlement such as limited memory. As participants were not allowed to rewatch previous demonstrations to enforce scaffolding order, similar demonstrations (in correspondingly similar environments) were sometimes mistaken to have shown different behaviors in the same environment.

Future work: Future iterations would benefit from "marking critical features" that "accentuates certain features of the task that are relevant", as suggested by Wood et al. (1976). For example, imagine showing two side-by-side demonstrations in the delivery domain, one where the robot exits because of the many mud patches in its path and one where the robot completes the delivery because of one fewer mud patch in its path. Outlining the presence and absence of the critical mud patch with a salient border in the two demonstrations respectively would help highlight the relevant cause for the change in robot behavior to the learner.

7.4 Testing

Objective and subjective results strongly support BEC area as a measure of test difficulty for human learners. Following studies may thus use tests of varying BEC areas and difficulties to evaluate and track the learner's understanding throughout the learning process.

Future work: Effective scaffolding is contingent on maintaining an accurate model of the learner's current abilities. Though this work assumed disjoint teaching and testing phases, learning is far more dynamic in reality. Future work should therefore explore how to select an initial set of tests that can accurately discern the learner's current knowledge, and also to know when to switch between teaching and testing throughout the learning process.

7.5 Real-world Applicability

Though the proposed method of machine teaching is theoretically general, there are additional considerations that must be addressed for real-world applicability.

First, a robot's policy may be a function of many parameters. Though performing IRL in a high-dimensional space may sometimes be warranted, humans naturally exhibit a bias toward simpler explanations with fewer causes (Lombrozo, 2016) and can only effectively reason about a few variables at once (e.g. Halford et al. (2005) suggest the limit to be around four). Thus, future work may examine approximating a high-dimensional policy with a low-dimensional policy that can be conveyed instead with minimal loss. Additionally, scaffolding methods that explicitly convey only a subset of the reward weights at a time should be developed as previously noted.

Second, a robot's entire trajectory will not always be necessary or reasonable to convey if it is lengthy. Thus techniques that extract and convey only the informative segments along with sufficient context will be important. For segments that are infeasible to convey in the real world (e.g due to necessary preconditions not being met), demonstrations may be given in simulation instead.

8 CONCLUSION

As robots continue to gain useful skills, their ability to teach them to humans will benefit those looking to acquire said skills and also facilitate fluent collaboration with humans. In this work, we thus explored how a robot may teach by providing demonstrations of its skill that are tailored for human learning.

We augmented the common model of humans as inverse reinforcement learners with insights from learning theory and cognitive science to better accommodate human learning. Scaffolding provided demonstrations that increase in informativeness and difficulty, aiming to ease the learner into the skill being taught. Furthermore, simple demonstrations that conveyed a discernible pattern were favored to minimize potentially misleading distractions and instead highlight critical features. Finally, a measure for quantifying the difficulty of tests was proposed toward effective evaluation of learning progress.

User studies strongly correlated our measure of test difficulty with human performance and confidence. Favoring simplicity and pattern discovery when selecting teaching demonstrations also led to a significant increase in performance for high difficulty tests. However, scaffolding failed to produce a significant effect on the test performance, informing both the shortcomings of the current implementation and the ways it can be improved in future iterations. Finally, though this work assumed disjoint teaching and testing phases with a static human model, effective scaffolding requires the teacher query, maintain, and leverage a dynamic model of the student to tailor the learning appropriately. We leave this as an exciting direction for future work.

DATA AVAILABILITY STATEMENT

The code for the human teaching techniques can be found in the following repository: https://github.com/SUCCESS-MURI/machine-teaching-human-IRL. The code for generating the user study (including videos of the teaching and testing demonstrations) and the data corresponding to our results can be found in the following repository: https://github.com/SUCCESS542 MURI/psiturk-machine-teaching.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of Carnegie Mellon University. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors equally contributed to the ideas presented in this paper, i.e. the techniques for human teaching and user studies design. ML implemented the techniques and user studies, ran the user studies, and analyzed the data. The manuscript was prepared, revised, and approved by all authors.

FUNDING

This work was supported by the Office of Naval Research award N00014-18-1-2503 and Defense Advanced Research Projects Agency (DARPA)/Army Research Office (ARO) award W911NF-20-1-0006. The views and conclusions contained in this document are of the authors and should not be interpreted as representing official policies, expressed or implied, of DARPA, ARO, or U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation herein.

ACKNOWLEDGMENTS

We would like to thank Vignesh Rajmohan and Meghna Behari for their assistance in creating the user study, and Pallavi Koppol for serving as an independent coder and for sharing her user study and data analysis templates.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.693050/full#supplementary-material

REFERENCES

Abbeel, P., and Ng, A. Y. (2004). Apprenticeship Learning via Inverse Reinforcement Learning. In Proceedings of the twenty-first international conference on Machine learning.

Abel, D. (2019). Simple Rl: Reproducible Reinforcement Learning in python. In ICLR Workshop on Reproducibility in Machine Learning.

Altman, D. G. (1990). *Practical Statistics for Medical Research* (Boca Raton, FL: CRC Press).

Amir, D., and Amir, O. (2018). Highlights: Summarizing Agent Behavior to People. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (Richland, South Carolina: International Foundation for Autonomous Agents and Multiagent Systems), 1168–1176.

Amir, O., Doshi-Velez, F., and Sarne, D. (2019). Summarizing Agent Strategies. *Auton. Agent Multi-agent Syst.* 33, 628–644. doi:10.1007/s10458-019-09418-w

Amitai, Y., and Amir, O. (2021). *"I Don't Think So": Disagreement-Based Policy Summaries for Comparing Agents* arXiv preprint arXiv:2102.03064.

Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action Understanding as Inverse Planning. *Cognition* 113, 329–349. doi:10.1016/j.cognition.2009.07.005

Baker, C., Saxe, R., and Tenenbaum, J. (2011). Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. In Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 33.

Brown, D. S., and Niekum, S. (2018). *Thirty-Second AAAI Conference on Artificial Intelligence*. Efficient Probabilistic Performance Bounds for Inverse Reinforcement Learning.

Brown, D. S., and Niekum, S. (2019). Machine Teaching for Inverse Reinforcement Learning: Algorithms and Applications. In Proceedings of the AAAI Conference on Artificial Intelligence Aaai 33, 7749–7758. doi:10.1609/aaai.v33i01.33017749

Cakmak, M., and Lopes, M. (2012). Algorithmic and Human Teaching of Sequential Decision Tasks. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. 1536–1542.

Collins, A., Brown, J. S., and Newman, S. E. (1988). Cognitive Apprenticeship. *Thinking: J. Philos. Child.* 8, 2–10. doi:10.5840/thinking19888129

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based Competition between Prefrontal and Dorsolateral Striatal Systems for Behavioral Control. *Nat. Neurosci.* 8, 1704–1711. doi:10.1038/nn1560

Dietterich, T. G. (1998). The Maxq Method for Hierarchical Reinforcement Learning. In Proceedings of the Fifteenth International Conference on Machine Learning. 118–126.

Griffiths, T. L. (2020). Understanding Human Intelligence through Human Limitations. *Trends Cogn. Sci.*

Guneysu Ozgur, A., Özgür, A., Asselborn, T., Johal, W., Yadollahi, E., Bruno, B., et al. (2020). Iterative Design and Evaluation of a Tangible Robot-Assisted Handwriting Activity for Special Education. *Front. Robot. AI* 7, 29. doi:10.3389/frobt.2020.00029

Halford, G. S., Baker, R., McCredden, J. E., and Bain, J. D. (2005). How many Variables Can Humans Process? *Psychol. Sci.* 16, 70–76. doi:10.1111/j.0956-7976.2005.00782.x

Huang, S. H., Bhatia, K., Abbeel, P., and Dragan, A. D. (2018). Establishing Appropriate Trust via Critical States. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 3929–3936.

Huang, S. H., Held, D., Abbeel, P., and Dragan, A. D. (2019). Enabling Robots to Communicate Their Objectives. *Auton. Robot* 43, 309–326. doi:10.1007/s10514-018-9771-0

Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends Cognitive Sciences* 20, 589–604. doi:10.1016/j.tics.2016.05.011

Jara-Ettinger, J. (2019). Theory of Mind as Inverse Reinforcement Learning. *Curr. Opin. Behav. Sci.* 29, 105–110. doi:10.1016/j.cobeha.2019.04.010

Jern, A., Lucas, C. G., and Kemp, C. (2017). People Learn Other People's Preferences through Inverse Decision-Making. *Cognition* 168, 46–64. doi:10.1016/j.cognition.2017.06.017

Koo, T. K., and Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropractic Med.* 15, 155–163. doi:10.1016/j.jcm.2016.02.012

Lage, I., Lifschitz, D., Doshi-Velez, F., and Amir, O. (2019). Exploring Computational User Models for Agent Policy Summarization. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19 (Red Hook, New York: International Joint Conferences on Artificial Intelligence Organization), 1401–1407. doi:10.24963/ijcai.2019/194

Landis, J. R., and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 159–174. doi:10.2307/2529310

Liu, J., and Zhu, X. (2016). The Teaching Dimension of Linear Learners. *J. Machine Learn. Res.* 17, 1–25.

Lombrozo, T. (2016). Explanatory Preferences Shape Learning and Inference. *Trends Cogn. Sci.* 20, 748–759. doi:10.1016/j.tics.2016.08.001

Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., et al. (2014). The Child as Econometrician: A Rational Model of Preference Understanding in Children. *PloS one* 9, e92160. doi:10.1371/journal.pone.0092160

McHugh, M. L. (2012). Interrater Reliability: the Kappa Statistic. *Biochem. Med.* 22, 276–282. doi:10.11613/bm.2012.031

Ng, A. Y., and Russell, S. (2000). Algorithms for Inverse Reinforcement Learning. In in Proc. 17th International Conf. on Machine Learning.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., Peters, J., et al. (2018). An Algorithmic Perspective on Imitation Learning. *FNT in Robotics* 7, 1–179. doi:10.1561/2300000053

Palan, S., and Schitter, C. (2018). Prolific.ac-A Subject Pool for Online Experiments. *J. Behav. Exp. Finance* 17, 22–27. doi:10.1016/j.jbef.2017.12.004

Paulraj, S., and Sumathi, P. (2010). A Comparative Study of Redundant Constraints Identification Methods in Linear Programming Problems. *Math. Probl. Eng.*, 2010.

Reiser, B. J. (2004). Scaffolding Complex Learning: The Mechanisms of Structuring and Problematizing Student Work. *J. Learn. Sci.* 13, 273–304. doi:10.1207/s15327809jls1303_2

Sampayo-Vargas, S., Cope, C. J., He, Z., and Byrne, G. J. (2013). The Effectiveness of Adaptive Difficulty Adjustments on Students' Motivation and Learning in an Educational Computer Game. *Comput. Edu.* 69, 452–462. doi:10.1016/j.compedu.2013.07.004

Sandygulova, A., Johal, W., Zhexenova, Z., Tleubayev, B., Zhanatkyzy, A., Turarova, A., et al. (2020). Cowriting Kazakh: Learning a New Script with a Robot. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. 113–120.

Shteingart, H., and Loewenstein, Y. (2014). Reinforcement Learning and Human Behavior. *Curr. Opin. Neurobiol.* 25, 93–98. doi:10.1016/j.conb.2013.12.004

Vul, E., Goodman, N., Griffiths, T. L., and Tenenbaum, J. B. (2014). One and Done? Optimal Decisions from Very Few Samples. *Cogn. Sci.* 38, 599–637. doi:10.1111/cogs.12101

Watkins, C. J. C. H., and Dayan, P. (1992). Q-learning. *Machine Learn.* 8, 279–292. doi:10.1023/a:1022676722315

Williams, J. J., Lombrozo, T., and Rehder, B. (2010). Why Does Explaining Help Learning? Insight from an Explanation Impairment Effect. In Proceedings of the Annual Meeting of the Cognitive Science Society. vol. 32.

Wood, D., Bruner, J. S., and Ross, G. (1976). The Role of Tutoring in Problem Solving. *J. Child. Psychol. Psychiat* 17, 89–100. doi:10.1111/j.1469-7610.1976.tb00381.x

Wunderlich, K., Dayan, P., and Dolan, R. J. (2012). Mapping Value Based Planning and Extensively Trained Choice in the Human Brain. *Nat. Neurosci.* 15, 786–791. doi:10.1038/nn.3068

Zhu, X., Singla, A., Zilles, S., and Rafferty, A. N. (2018). *An Overview of Machine Teaching* (arXiv preprint arXiv:1801.05927.

Zhu, X. (2015). Machine Teaching: an Inverse Problem to Machine Learning and an Approach toward Optimal Education. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 4083–4087.

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 in Robotics and AI

ORIGINAL RESEARCH
published: 30 June 2021
doi: 10.3389/frobt.2021.679893

When Even a Robot Tutor Zooms: A Study of Embodiment, Attitudes, and Impressions

Junko Kanero[1]*, *Elif Tutku Tunalı*[1], *Cansu Oranç*[2], *Tilbe Göksun*[3] *and Aylin C. Küntay*[3]

[1]*Faculty of Arts and Social Sciences, Sabancı University, Istanbul, Turkey,* [2]*MPRG iSearch, Max Planck Institute for Human Development, Berlin, Germany,* [3]*Department of Psychology, Koç University, Istanbul, Turkey*

This study used an online second language (L2) vocabulary lesson to evaluate whether the physical body (i.e., embodiment) of a robot tutor has an impact on how the learner learns from the robot. In addition, we tested how individual differences in attitudes toward robots, first impressions of the robot, anxiety in learning L2, and personality traits may be related to L2 vocabulary learning. One hundred Turkish-speaking young adults were taught eight English words in a one-on-one Zoom session either with a NAO robot tutor (N = 50) or with a voice-only tutor (N = 50). The findings showed that participants learned the vocabulary equally well from the robot and voice tutors, indicating that the physical embodiment of the robot did not change learning gains in a short vocabulary lesson. Further, negative attitudes toward robots had negative effects on learning for participants in the robot tutor condition, but first impressions did not predict vocabulary learning in either of the two conditions. L2 anxiety, on the other hand, negatively predicted learning outcomes in both conditions. We also report that attitudes toward robots and the impressions of the robot tutor remained unchanged before and after the lesson. As one of the first to examine the effectiveness of robots as an online lecturer, this study presents an example of comparable learning outcomes regardless of physical embodiment.

Keywords: human-robot interaction, second language learning (L2 learning), embodiment, attitudes, impressions

OPEN ACCESS

Edited by:
Wafa Johal,
University of New South Wales, Australia

Reviewed by:
Pauline Chevalier,
Italian Institute of Technology (IIT), Italy
Hatice Kose,
Istanbul Technical University, Turkey

***Correspondence:**
Junko Kanero
jkanero@sabanciuniv.edu

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: *12 March 2021*
Accepted: *25 May 2021*
Published: *30 June 2021*

Citation:
Kanero J, Tunalı ET, Oranç C, Göksun T and Küntay AC (2021) When Even a Robot Tutor Zooms: A Study of Embodiment, Attitudes, and Impressions. Front. Robot. AI 8:679893. doi: 10.3389/frobt.2021.679893

INTRODUCTION

Social robots, robots that interact and communicate with humans by following the behavioral norms of human-human interactions (e.g., Bartneck and Forlizzi, 2004; Kanero et al., 2018), are becoming abundant across a variety of settings such as homes, hospitals, and schools. A particularly interesting application of social robots is language education because of the significance of the topic as well as the unique characteristics of social robots. Language education is critical for people of all ages. For children, language abilities are known to predict future academic achievement and social skills (Hoff, 2013; Milligan et al. 2007); for adults, language skills can broaden social and occupational opportunities (e.g., Paolo and Tansel, 2015). Learning another language can also contribute to the development of cognitive skills in children (Kovács and Mehler, 2009), and the attainment of them in older adults (Bialystok et al., 2004). Importantly, a wealth of research in psychology and education suggests that learning both first (L1) and second language (L2) requires *interactions* (Verga and Kotz, 2013; Konishi et al., 2014; Lytle and Kuhl, 2017). As a social agent with a physical body, a robot can play the role of a tutor through vocal, gestural, and facial expressions to provide an interactive learning experience (Han et al., 2008; Kennedy et al., 2015; Kanero et al., 2018). The current study focuses on *embodiment* and examines whether and how important it is for L2 learners to interact with a robot tutor with a physical body.

The general bodily affordances of social robots were suggested to improve the learning experience as they can engage with the learners' physical world and elicit social behaviors from them (Belpaeme et al., 2018). For instance, when teaching a new word, robots can perform gestures with their hands to depict the target object or direct the learner's attention to the object with their eyes, both of which are an integral part of interacting and learning with robots (Admoni and Scassellati, 2017; Kanero et al., 2018). Some studies indicate that interacting with a robot in person or through a screen may not have much of a difference in terms of learning (e.g., Kennedy et al., 2015), and studies on language learning with intelligent virtual agents provide support to this (Macedonia et al., 2014). In fact, a study on second language learning found participants performing worse after interacting with a physically present robot as opposed to its virtual version or a voice-only tutor, speculatively because it was too novel and interesting, hence distracting, for the participants (Rosenthal-von der Pütten et al., 2016). On the other hand, there is also research suggesting that interacting with a physically present robot may yield better outcomes. For instance, one study found that adults performed better in solving logic puzzles when they were partnered off with a physically present robot as opposed to a disembodied voice or a video of a robot (Leyzberg et al., 2012), though solving a logic puzzle is inherently different from learning a language.

Embodiment has been defined in many different ways partially because the term is used in various disciplines including philosophy, psychology, computer science, and robotics (see Deng et al., 2019). One of the clear definitions provided by roboticists is that of Pfeifer and Scheier (1999): "In artificial systems, the term refers to the fact that a particular agent is realized as a physical robot or as a simulated agent" (p. 649). Focusing on the social aspect, Barsalou et al. (2003) states that embodiment is the "states of the body, such as postures, arm movements, and facial expressions, [which] arise during social interaction and play central roles in social information processing" (p. 43). In human-robot interaction, Li (2015) made a distinction between what he calls *physical presence* and *physical embodiment* to systematically evaluate the different bodily affordances of robots. According to Li (2015), *physical presence* differentiates a robot in the same room with the user and a robot appearing on the screen. On the other hand, *physical embodiment* differentiates a (co-present or telepresent) materialized robot and a virtual agent (e.g., a computer-generated image of a robot).

The review by Li (2015) concluded that the physical presence of the robot, but not its embodiment, has a positive influence on social interactions. Critically, however, the conclusion was drawn based on four studies from three publications only. Overall, while previous research provides valuable insights into how different dimensions of physicality influence human-robot interaction, they fall short in revealing the difference between having and not having a body and face on learning outcomes. Although their appearance can simulate different animate agents such as a human or an animal, all social robots have a body and face. How does this influence people's learning, as opposed to not having either? Following the distinctions drawn by Li (2015), we compare a robot tutor (embodied but not physically present) with a voice-only tutor (not embodied nor physically present) in an online lesson to understand the effects of physical embodiment.

Research also suggests that embodiment may have different implications for different people, as in individuals with Autism Spectrum Disorder struggling with understanding the emotions of a virtual agent than a real agent, whether it is a robot or a human, in contrast to typically developed individuals (Chevalier et al., 2017). People's varying attitudes toward robots may also influence their preference for a physical or virtual robot (Ligthart and Truong, 2015). Another study with children also suggests that age and experience may diminish the effect of physical presence, as it found that younger children with hearing impairments learned more words in sign language when they interacted with a physically present robot than a video of it, whereas older children with more experience in sign language equally benefited from both (Köse et al., 2015). Therefore, the current study further explores interrelations among individual differences (specifically attitudes toward robots, first impressions of the robot tutor, anxiety about learning a second language, and personality traits) and learning outcomes across different degrees of embodiment.

Although not much is known specifically about the effects of individual differences in learning with robots, some studies have explored how attitudes and personality are related to the ways in which a person interacts with a robot. For example, the patterns of speech and eye gaze were observed while adults built an object with a humanoid robot (Ivaldi et al., 2017). The study found that individuals with negative attitudes toward robots tended to look less at the robot's face and more at the robot's hands. In another study, when approached by a robot, individuals with high levels of negative attitudes toward robots and the personality trait of neuroticism kept a larger personal space between the robot and themselves (Takayama and Pantofaru, 2009).

In the case of language learning, Kanero et al. (2021) were first to examine how attitudes toward robots, anxiety about learning L2, and personality may predict the learning outcomes of a robot-led L2 vocabulary lesson. The study found that negative attitudes measured through the Negative Attitudes toward Robots Scale (NARS; Nomura et al., 2006) as well as anxiety about learning L2 measured through the Foreign Language Classroom Anxiety Scale (FLCAS; Horwitz et al., 1986) predicted the number of words participants learned in an in-person vocabulary lesson with a robot tutor. The results also showed that the robot was an effective language tutor, akin to a human tutor. However, it is unclear whether the tutor robot is as effective when it is not physically present, and whether individual differences such as attitudes toward robots and L2 anxiety predict the learning outcomes for a telepresent robot tutor.

In addition to the individual difference measures used in the previous study (Kanero et al., 2021), the current study also assesses the learners' impressions of robots, which are expected to affect their engagement in the long run. Previous studies in human-human interaction suggest that the first impression is formed very quickly after just seeing a picture of an individual and might remain unchanged even after meeting and interacting with the same individual in person (e.g., Gunaydin et al., 2017; see also Willis and Todorov, 2006). However, it is unclear whether the same principle applies to commercial social robots (e.g., NAO), which are inanimate objects with a homogeneous appearance shared across individuals. Therefore, in the current study, we included an additional measure to examine if the impressions of the robot have a role in robot-led learning. Further, we evaluate whether the impressions of the robot tutor as well as attitudes toward robots change before and after interacting with the robot tutor.

In summary, this study explores the impact of having a body in robot-led language lessons by comparing a robot tutor and a voice-only tutor in terms of learning outcomes as well as the influence of attitudes, impressions, L2 anxiety, personality. We also report the details of the learner's general attitudes toward roborts, impressions of the robot tutor, and preferences to the specific type of tutor (robot vs. voice vs. human). In the Discussion, we also compare our data to the data of the previous study (Kanero et al., 2021) to address whether the physical presence in robot-led language lessons would affect these factors.

METHODS

Participants

The dataset consisted of 100 native Turkish-speaking young adults: 50 in the robot tutor condition (*age range* = 18–32 years; M_{age} = 23.49 years; *SD* = 2.53; 33 females, 17 males), and 50 in the voice tutor condition (*age range* = 18–35 years; M_{age} = 24.15 years; *SD* = 3.62; 33 females, 16 males, 1 other). We relied on a convenience sample, and participants were recruited through advertisements on social media as well as word of mouth. Before the lesson, the average English test score of participants (Quick Placement Test; University of Cambridge Local Examinations Syndicate [UCLES], 2001) was 39.68 out of 60 in the robot tutor condition (*score range* = 16–58; *SD* = 9.07) and 37.64 in the voice tutor condition (*score range* = 20–55; *SD* = 9.25). Participants had no known vision or hearing impairments. One participant in the robot tutor condition did not show up for the second session, and thus the two delayed language tests and the post-lesson survey were not administered to this participant. In addition, one participant in the robot tutor condition was not taught one of the eight vocabulary words due to a technical error, and thus the test data for that word were not used. Participants were given a gift card for their participation.

Materials and Procedures

The experiment was completed via the online video call software *Zoom* (https://zoom.us) in two sessions. In the first session, participants first filled out a demographic form. They then completed a short English language test (Quick Placement Test; UCLES, 2001), and a questionnaire assessing their attitudes toward robots, L2 anxiety, personality traits, and their impression of the robot or voice tutor. The test and questionnaires were administered using the online survey platform Qualtrics (https://www.qualtrics.com). Then, participants received a one-on-one English lesson either from the robot or the voice tutor. For the lesson, participants were sent to a breakout room[1], and participants were alone with the tutor. Immediately after the lesson, participants in both conditions completed two measures of learning (i.e., immediate production and receptive tests). The second session took place one week later, and participants connected via Zoom again and completed the same vocabulary tests (i.e., delayed production and receptive tests). The same set of tests and surveys were administered in the robot tutor and voice tutor conditions, but in the voice tutor condition, the term "voice assistant" was used in place of "robot" for the surveys on the attitudes, impressions, and preference (see **Figure 1** for a schematic representation of the procedure, and **Figure 2** for the appearance of the robot and voice-only tutors).

Negative Attitudes Toward Robots

Negative Attitudes toward Robots Scale (NARS; Nomura et al., 2006) was used to assess attitudes toward robots. The NARS consists of 14 questions divided into three subordinate scales: negative attitude toward interacting with robots (S1), negative attitude toward the social influence of robots (S2), and negative attitude toward emotions involved in the interaction with robots (S3). The Turkish version of the NARS (Kanero et al., 2021) was used. Participants rated how well each of the statements represented their attitudes toward robots on a scale of 1–5 (1: I strongly disagree/Kesinlikle katılmıyorum, 2: I disagree/ Katılmıyorum, 3: Undecided/Kararsızım, 4: I agree/ Katılıyorum, 5: I strongly agree/Kesinlikle katılıyorum). In the voice tutor condition, the word "robot" on the NARS scale was replaced by "voice assistant."

Impressions of the Robot Tutor

To assess participants' impressions of the robot/voice tutor, we administered an impression survey with 17 questions used by Gunaydin and her colleagues (2017; available publicly at https://osf.io/nhmtw/?view_only=9f6efafeba4b48dc9b6a73b6a3d145ee). The survey shows a photograph of the robot or voice tutor, depending on the condition, and consists of two parts: The first eight questions ask participants to rate their willingness to engage and interact with the target in the future (e.g., This robot/voice assistant seems like a robot/voice assistant I would like to get to know/Tanımak istediğim bir robot/sesli asistan gibi gözüküyor) on a scale of 1–7 (1: I fully disagree/Hiç katılmıyorum, 2: I disagree/Katılmıyorum, 3: I somewhat disagree/Kısmen katılmıyorum, 4: I neither agree nor disagree/Ne katılıyorum ne de katılmıyorum, 5: I somewhat agree/ Kısmen katılıyorum, 6: I agree/katılıyorum, 7: I fully agree/Tamamen katılıyorum). The next nine questions ask participants to rate how

[1]Breakout room is a feature in Zoom that allows the host to split one Zoom session into multiple separate subsessions whereby participants in separate breakout rooms do not see each other. We put the participant into a separate breakout room away from the Experimenter and the tutor so that participants do not need to feel watched or pressured.

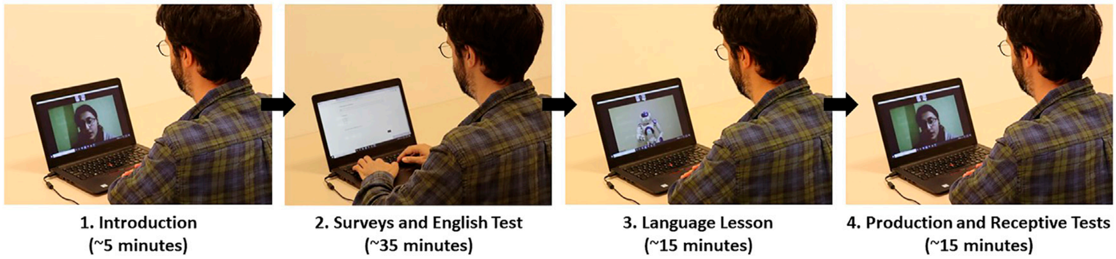


FIGURE 1 | The procedure of the lesson from the participant's perspective. In the voice-only tutor condition, the voice sound spectrum appeared instead of the robot (see **Figure 2** and **Supplementary Material**). Step 4 (Production and Receptive Tests) was repeated one week later.

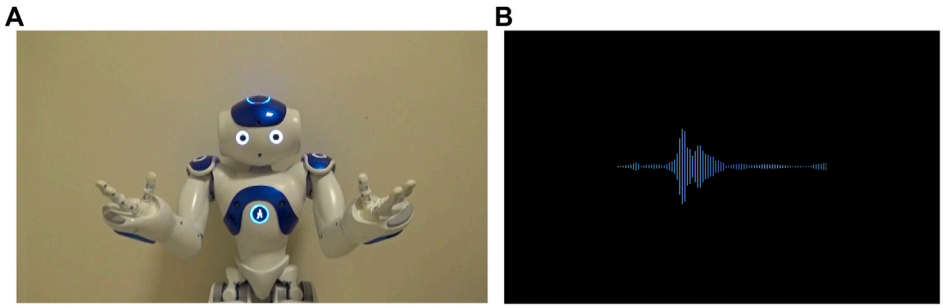


FIGURE 2 | The appearance of the robot tutor **(A)** and the voice tutor **(B)**. See **Supplementary Material** for the videos of the robot and voice tutors).

their interaction will be with the robot/voice assistant (e.g., How much do you think you will like this robot/voice assistant?/Bu robotu/sesli asistanı ne kadar seveceğinizi düşünüyorsunuz?) on a scale of 1–7 (1: Not at all/Hiç, 7: Very much/Çok fazla). After the lesson, participants rated the same items but were told to rate the statements based on their interactions with their tutor. The original survey in English was translated into Turkish by the second author and research assistants who are native speakers of Turkish. To adapt to our study, the word "person" was replaced with "robot" for the robot tutor condition and "voice assistant" for the voice tutor condition.

L2 Anxiety

The Turkish version of the Foreign Language Classroom Anxiety Scale (FLCAS; Horwitz et al., 1986) translated by Aydın et al. (2016) was administered. The FLCAS consists of 33 statements (e.g., I never feel quite sure of myself when I am speaking in my foreign language class/Yabancı dil derslerinde konuşurken kendimden asla emin olamıyorum.) to be rated on a scale of 1–5 (1: I fully disagree/Hiç katılmıyorum, 2: I disagree/Katılmıyorum, 3: I neither agree nor disagree/Ne katılıyorum ne de katılmıyorum, 4: I agree/Katılıyorum, 5: I fully agree/Tamamen katılıyorum).

Personality Traits

The Turkish version of a personality inventory was used to test the five personality traits – openness to experience, conscientiousness, extroversion, agreeableness, and neuroticism (Demir and Kumkale, 2013). This survey included 44 questions addressing each of the five traits – 7 items for conscientiousness (e.g., I stick to my plans/Yaptığım planlara sadık kalırım); 10 items for neuroticism (e.g., I am depressed/Depresifimdir); 9 items for each of openness to experience (e.g., My interests are very diverse/İlgi alanlarım çok çeşitlidir), extroversion (e.g., I am talkative/Konuşkanımdır), and agreeableness (e.g., I am helpful/Yardımseverimdir). Participants rated how well each of the statements represented their personality on a scale of 1–5 (1: I strongly disagree/Kesinlikle katılmıyorum, 2: I disagree/Katılmıyorum, 3: I neither agree nor disagree/Ne katılıyorum, ne de katılmıyorum, 4: I agree/Katılıyorum, 5: I strongly agree/Kesinlikle katılıyorum).

Post-Lesson Vocabulary Tests

Immediately after the lesson, we first administered the production vocabulary test (hereafter the immediate production test), and then the receptive vocabulary test (hereafter the immediate receptive test). To assess to what extent vocabulary was retained over time, participants completed the same measures again after a delay of one week (delayed post-lesson tests). The definitions of the target words used in the production test were the same as the definitions used in the lesson. In the receptive test, the pictures from the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn and Dunn, 2007), which correspond to the target words, were used. In the production test, the experimenter provided the definitions of the learned English words one by one in a randomized order, and the participant was asked to say the corresponding English word. In the receptive test, the participant heard the learned English word and was asked to choose a picture that matched the word from four options. The delayed post-lesson tests were conducted via Zoom seven days

TABLE 1 | The target words and their definitions used in the study.

Word	Definition
Upholstery	*Bu kelime döşemelik kumaş anlamına gelir* (This word means fabric used to make a soft covering)
barb	*Bu kelime çengel ya da kanca anlamına gelir* (This word means the tip of an arrow or fishhook)
Angler	*Bu kelime olta ile balık tutan kimse anlamına gelir* (This word means a person who fishes with hook and line)
Caster	*Bu kelime bir şeye takılan küçük tekerlek anlamına gelir* (This word means a little wheel attached to something)
Dromedary	*Bu kelime tek hörgüçlü deve anlamına gelir* (This word means a one-humped camel)
Cairn	*Bu kelime taş yığını anlamına gelir* (This word means a mound of stones)
Derrick	*Bu kelime petrol kuyusu üzerindeki kule anlamına gelir* (This word means a tower over an oil well)
Cupola	*Bu kelime bir çatı üstüne inşa edilen küçük kubbe benzeri yapı anlamına gelir* (This word means a rounded vault-like structure built on top of a roof)

after the lesson. Due to schedule conflicts, however, three participants in the robot tutor condition and two participants in the voice tutor condition completed these tests after six days, while four participants in each condition completed the tests after eight days. Also, three participants in the voice tutor condition completed the test after nine days.

Tutor Preference

After the delayed post-lesson tests, we also asked participants to rate how much they want to learn English from a human, a robot, and a voice assistant. A scale of 1–5 was used (1: I certainly do not want/Kesinlikle istemem, 2: I do not want/İstemem, 3: I neither want nor not want/Ne isterim ne istemem, 4: I want/İsterim, 5: I certainly want/Kesinlikle isterim).

English Lesson With the Robot or Voice Tutor

Following the previous study (Kanero et al., 2021), participants were taught eight English nouns – upholstery, barb, angler, caster, dromedary, cairn, derrick, and cupola (see **Table 1**; see Kanero et al. (2021) for the details of the word selection process).

In both tutor conditions, the robot or voice tutor briefly chatted with the participant and explained the structure of the lesson first, and then introduced the words one by one. Each target word was taught in four steps:

1) The tutor introduced the target L2 (English) word and asked the participant whether the participant already knew the word (Note that none of the participants knew any of the words).
2) The tutor introduced the definition of the target word in L1 (Turkish, see **Table 1**).
3) The tutor asked the participant to utter the target word following the tutor, three times.
4) The tutor again defined the word and asked the participant to repeat the definition.

After learning every two target words, the participant was given a mini quiz in which the tutor provided the definitions of the target words and asked the participant for the corresponding word. The lesson lasted for about 15 min. At the end of the lesson, the robot or the voice tutor asked the participant to return to the previous room and find the experimenter they met prior to the lesson. The human experimenter administered the immediate production and receptive vocabulary tests.

To use the same voice in English and Turkish speech, we recorded the speech of a female bilingual experimenter and added sound effects to make the speech sound robot-like. The same set of speech sounds were used for both the robot and voice tutors. The visuals of both tutors were presented as a series of seamlessly transitioning videos on Zoom. The movements of the robot tutor were filmed (see **Figure 2A**), whereas the soundwaves of the voice tutor were created using Adobe After Effect (https://www.adobe.com/products/aftereffects.html; **Figure 2B**).

The robot tutor provided no facial expressions but moved its head and arms during the lesson to keep the participant engaged. Most actions were chosen from the Animated Speech library of SoftBank Robotics (http://doc.aldebaran.com/2-1/naoqi/audio/alanimatedspeech_advanced.html), although some were created by the first author to better suit the lesson.[2] While pronouncing the target L2 word and its definition, the robot stood still without any movements to avoid the motor sound of the robot hindering the hearing. There were unavoidable behavioral differences between the two tutors (e.g., the motor sound of the robot), but otherwise, the differences between the two tutors were kept minimal.

RESULTS

Robot vs. Voice Tutor

We first examined if participants in the robot tutor and voice tutor conditions differed in their post-lesson test scores. We compared the two tutor conditions across all four learning outcome measures: immediate production test, immediate receptive test, delayed production test, and delayed receptive test. We conducted simple Generalized Linear Mixed Models (GLMMs) on each post-lesson test with Tutor Type (robot vs. voice) as a fixed effect and Word as random intercepts.[3] In this model, we also added the pre-lesson English test scores as an additional fixed effect to control for the difference in English proficiency between the conditions. As shown in **Figure 3**, participants did not differ in terms of learning outcomes across conditions (immediate production test, $B = 0.01$, $SE = 0.16$, $Z = 0.04$, $p = 0.968$; delayed production test, $B = -0.29$, $SE = 0.20$, $Z = -1.44$, $p = 0.149$; immediate receptive test, $B = 0.04$, $SE = 0.18$, $Z = 0.20$, $p =$

[2]The gestures used in the lesson were mostly generic except that, when the participant repeated the target word following the robot tutor, the robot made the "pinched fingers" gesture where all fingers were put together with the palm side up and the hand was moved up and down. This conventional gesture means "very good" in the Turkish culture.

[3]We used GLMMs in these analyses, because our data are not normally distributed, and because they allow us to analyze the responses of participants without averaging across trials (Jaeger, 2008). As the outcome (the scores of the four post-lesson tests) was a binary variable (correct vs. incorrect), logit (log-odds) was used as the link function. GLMMs were generated in R (R Development Core Team, 2016) using the *lme4.glmer* function (Bates, 2015). In all models, we included the random effect of item (e.g., L2 words) as some L2 vocabulary words may be inherently more difficult to learn than others. All models were fit by maximum likelihood using adaptive Gauss-Hermite quadrature (nAGQ = 1).

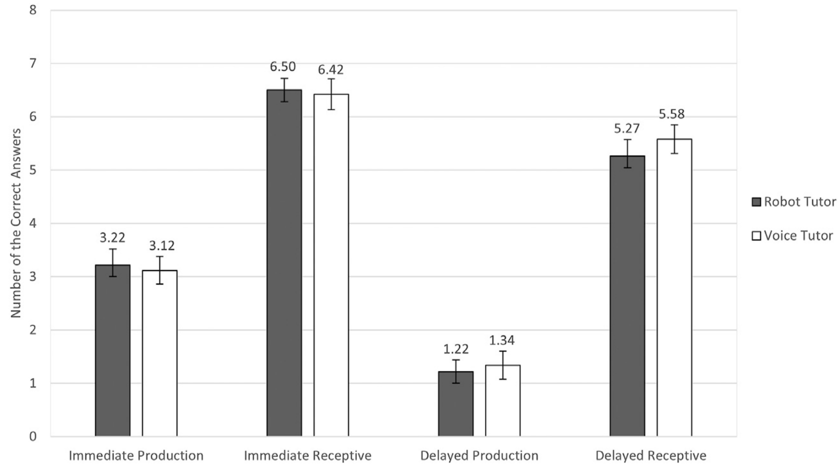


FIGURE 3 | Mean number of correct answers in the robot tutor and voice tutor conditions in the four post-lesson tests. *N* = 100 for the immediate production and receptive tests; *N* = 99 for the delayed production and receptive tests. The highest possible score for each test was eight. The error bars indicate the standard errors.

TABLE 2 | Descriptive statistics for the individual difference measures.

	Robot tutor			Voice tutor	
	α	*Mean*	*SD*	*Mean*	*SD*
NARS (14)	0.88	2.71	0.71	2.55	0.63
L2 anxiety (33)	0.95	2.57	0.79	2.48	0.69
Personality (44)					
Openness (9)	0.76	4.13	0.55	4.15	0.47
Conscientiousness (7)	0.81	3.20	0.74	3.22	0.59
Extroversion (9)	0.92	3.64	0.87	3.86	0.82
Agreeableness (9)	0.77	3.94	0.52	3.85	0.58
Neuroticism (10)	0.86	3.37	0.75	3.35	0.71
Impression (17)	0.92	4.19	1.27	3.51	1.02

N = 100. The number in parenthesis indicates the number of items in the scale.

0.845; or the delayed receptive test, B = -0.26, SE = 0.16, Z = -1.64, p = 0.101).

Predicting the Learning Outcomes With Individual Difference Factors

Next, we examined whether some participants learned better or worse from robots depending on their attitudes toward robots, the first impression of the robot or voice tutor, anxiety in L2 learning, and personality traits. As indicated by Cronbach's alphas in **Table 2**, each of these variables was measured reliably. Therefore, items measuring each construct were averaged to create relevant indices. For NARS, L2 anxiety, and personality, values ranged between 1 and 5. Higher values for NARS indicated having more negative attitudes toward robots; similarly, higher values for L2 Anxiety indicated having greater anxiety. For the impression survey, the values ranged between 1 and 7 and higher values indicated a more positive first impression.

Negative Attitudes Toward Robots

We built four separate GLMMs, one for each post-lesson test (immediate production, immediate receptive, delayed production, and delayed receptive), with Word as a random intercept to examine whether negative attitudes toward robots and voice assistants predicted the number of words participants learned. As shown in **Table 3**, in line with the previous study (Kanero et al., 2021), negative attitudes toward robots predicted the learning outcomes in a robot-led vocabulary lesson, though only in the delayed tests. Negative attitudes did not predict learning in the voice tutor condition.

First Impressions of the Robot

To evaluate the relation between the first impressions of the tutor and the learning outcomes, we followed the same steps and built GLMMs separately for the two tutor conditions. As shown in **Table 4**, there was no significant relation between learning outcomes and first impression in either condition.

L2 Anxiety and Personality Traits

The influence of L2 learning anxiety was similarly examined by building a GLMM for each post-lesson test for the robot tutor and voice tutor conditions with Word as a random intercept. In the robot tutor condition, L2 Anxiety predicted the scores of most tests except the immediate receptive test; in the voice tutor condition, the significance was found in the delayed production and receptive tests (**Table 5**).

We also built four GLMMs for each post-lesson test to evaluate the relevance of personality traits. In concert with the previous study (Kanero et al., 2021), the personality traits were not reliable predictors of the learning outcomes of the robot-led L2 lesson. In the robot tutor condition, extroversion was positively correlated with the immediate receptive test scores (B = 0.41, SE = 0.17, Z = 2.35, p = 0.019), and agreeableness was positively correlated with

TABLE 3 | GLMMs with NARS as the sole predictor for the four post-lesson scores.

	Robot tutor				Voice tutor			
	B	*SE*	*Z*	*p*	*B*	*SE*	*Z*	*p*
Immediate production	-0.08	0.15	-0.51	0.610	0.03	0.18	0.20	0.843
Immediate receptive	-0.20	0.18	-1.14	0.253	0.01	0.20	0.07	0.947
Delayed production	-0.45	0.22	-2.01	0.045	-0.03	0.21	-0.13	0.895
Delayed receptive	-0.31	0.15	-2.08	0.038	0.05	0.18	0.29	0.771

For the immediate tests, N *= 50 in both conditions; For the delayed tests,* N *= 49 in the robot tutor condition and* N *= 50 in the voice tutor condition.*

TABLE 4 | GLMMs with the first impression as the sole predictor for the four post-lesson scores.

	Robot tutor				Voice tutor			
	B	*SE*	*Z*	*p*	*B*	*SE*	*Z*	*p*
Immediate production	-0.06	0.09	-0.68	0.500	0.04	0.11	0.37	0.715
Immediate receptive	0.04	0.10	0.42	0.678	0.05	0.13	0.43	0.664
Delayed production	-0.03	0.11	-0.26	0.797	0.01	0.13	0.05	0.959
Delayed receptive	0.02	0.08	0.24	0.811	-0.04	0.11	-0.33	0.738

For the immediate tests, N *= 50 in both conditions; For the delayed tests,* N *= 49 in the robot tutor condition and* N *= 50 in the voice tutor condition.*

TABLE 5 | GLMMs with L2 Anxiety as the sole predictor for the four post-lesson scores.

	Robot tutor condition				Voice tutor condition			
	B	*SE*	*Z*	*p*	*B*	*SE*	*Z*	*p*
Immediate production	-0.43	0.14	-2.97	0.003	-0.16	0.16	-1.00	0.315
Immediate receptive	-0.22	0.16	-1.35	0.176	-0.15	0.18	-0.83	0.404
Delayed production	-0.42	0.20	-2.08	0.037	-1.36	0.27	-5.04	<0.001
Delayed receptive	-0.31	0.14	-2.25	0.025	-0.41	0.16	-2.53	0.012

For the immediate tests, N *= 50 in both conditions; For the delayed tests,* N *= 49 in the robot tutor condition and* N *= 50 in the voice tutor condition.*

the delayed receptive scores ($B = 0.65$, $SE = 0.23$, $Z = 2.77$, $p = 0.006$).

Attitudes, Impressions, and Preferences

Attitudes Toward Robots

With the purpose of assessing the change in attitudes after the interaction with the robot or voice tutor, the normality assumption of the data was first examined. In comparing the attitude scores between the two tutor conditions or between the pre- and post-lesson surveys, we performed a Shapiro-Wilk's test of normality. We then used *t*-tests when the two compared data are both normally distributed, and Wilcoxon Signed-Ranks Tests when the normality assumption was violated. The difference between the tutor conditions was not significant in either before ($Z = 0.97$, $p = 0.334$) nor after the lesson [t (97) = 1.17, $p = 0.244$]. Negative attitudes toward robots/voice assistants did not change before and after the lesson in the robot tutor condition ($Z = 1.10$, $p = 0.267$), nor the voice tutor condition [t (49) = 1.65, $p = 0.105$]. In other words, interacting with the tutor did not improve learners' attitudes toward the specific tutor (**Figure 4**).

Impressions of the Robot Tutor

A paired sample *t*-test on the impression survey indicated that, in the robot condition, participants' impressions of the robot before and after the lesson did not significantly change [t (48) = -0.22, $p = 0.407$]. In the voice tutor condition, on the other hand, the ratings were significantly higher after than before the lesson [t (49) = -3.78, $p < 0.001$]. In addition, independent paired *t*-tests demonstrated, that whereas the difference in the pre-lesson impression scores between the two tutor conditions was significant [t (98) = 2.89, $p = 0.005$], the two did not differ significantly in the post-lesson impression scores [t (97) = -0.06, $p = 0.954$]. These results indicate that, although the expectation was different for the two tutors, the impressions became comparable after having an actual interaction (see **Figure 5**).

Preference of Tutors

Wilcoxon Signed-Ranks tests suggest that participants in the robot tutor condition preferred a human tutor to a robot ($Z = 5.30$, $p < 0.001$) or a voice tutor ($Z = 5.52$, $p < 0.001$), but did not differ in their preference for a robot tutor and a voice tutor ($Z = 1.07$, $p = 0.286$; see **Figure 6**). In the voice tutor condition, participants also preferred a human tutor to a robot tutor ($Z = 5.59$, $p < 0.001$), and to a voice tutor ($Z = 5.39$, $p < 0.001$); and they also preferred a robot tutor to a voice tutor ($Z = 2.15$, $p = 0.031$). Participants in both tutor conditions did not significantly differ in their preference for human tutor ($Z = -1.27$, $p = 0.206$), robot tutor ($Z = -0.85$, $p = 0.397$) or voice tutor ($Z = -0.28$, $p = 0.778$).

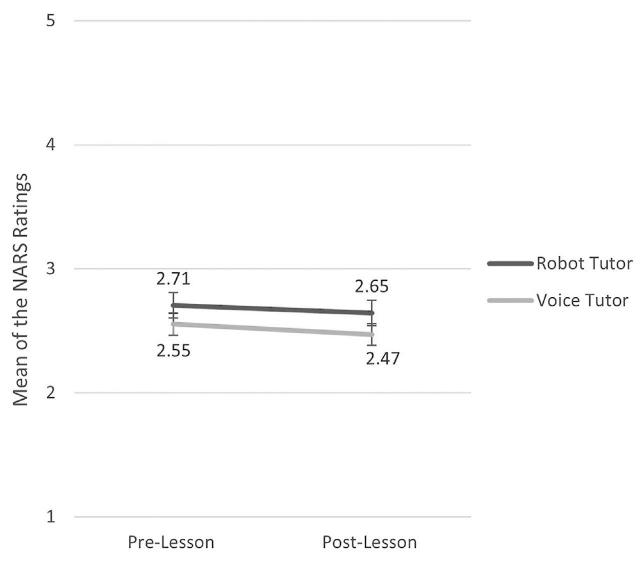


FIGURE 4 | Mean of the NARS ratings in the robot tutor and voice tutor conditions before and after the lesson. N = 100 for the pre-lesson NARS; N = 99 for the post-lesson NARS. The highest possible score for each test was 5. The error bars indicate the standard errors.

DISCUSSION

As the presence of social robots in our lives is becoming more and more prominent, it is critical to understand when and for whom robots can provide the most benefit. The present study examined the physical embodiment of robots and individual differences among learners to evaluate the effectiveness of robot tutors in an online L2 vocabulary lesson. To further understand the circumstances in which the robot tutor is effective, we also assessed how learners' individual differences in attitudes toward robots, impressions of the robot tutor, anxiety about L2 learning, and personality traits were related to their learning outcomes. Through a stringent evaluation using two different outcome measures at two time points, we found that embodiment did not affect learning in our lesson, and individuals with negative attitudes toward robots and L2 learning anxiety learned fewer words in the robot-led lesson.

Embodiment of the Robot Tutor

The learning outcomes were comparable on all four measures between the robot tutor and voice tutor conditions, and thus we did not see an advantage of the robot tutor having a body. Our results are in concert with previous research that did not find benefits of physical embodiment in learning (e.g., Kennedy et al., 2015). To further confirm the conclusion, we also conducted an exploratory analysis comparing the current data with the data from the in-person robot lesson in our previous study (Kanero et al., 2021). We built four GLMMs for each post-lesson test examining the main effects of 1) embodiment (in-person and Zoom robot tutors vs. Zoom voice tutor), and 2) physical presence (in-person robot tutor vs. Zoom robot and voice tutors). Neither embodiment nor physical presence was identified as a significant predictor (all p's > 0.080). Therefore, we found no evidence of the robot's embodiment or physical presence affecting the learning outcomes of the simple L2 vocabulary lesson. As discussed further in *Changes in Attitudes, Impressions, and Preferences From Before to After the Lesson* section, we did not find an impact of physical embodiment on the learning outcomes or impressions of the robot tutor after the lesson either. The context of our paradigm must be taken into consideration in interpreting these results, as our vocabulary learning task was solely conversational and did not require the robot to interact with the physical world. Embodiment

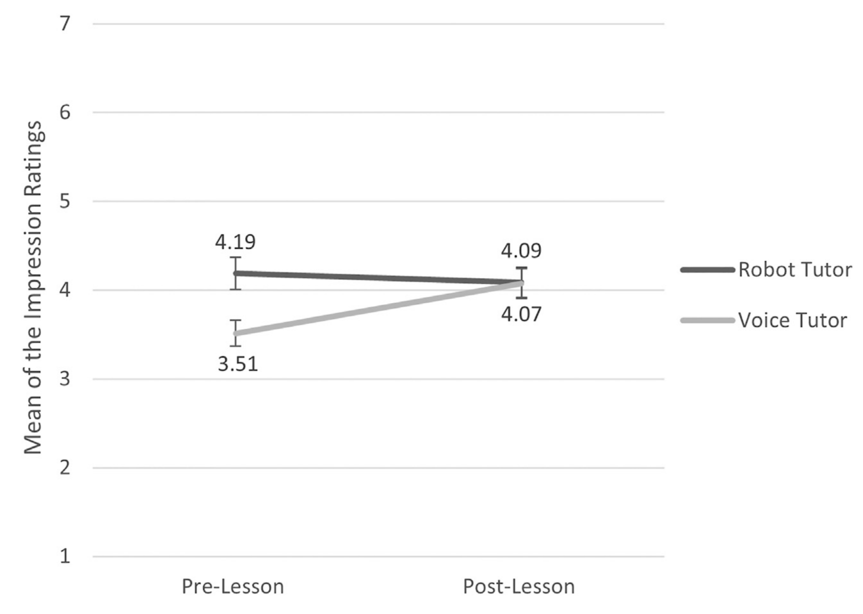


FIGURE 5 | Mean ratings of the impression survey in the robot tutor and voice tutor conditions before and after the lesson. N = 100 for pre-lesson and N = 99 for post-lesson. The highest possible score for each test was 7. The error bars indicate the standard errors.

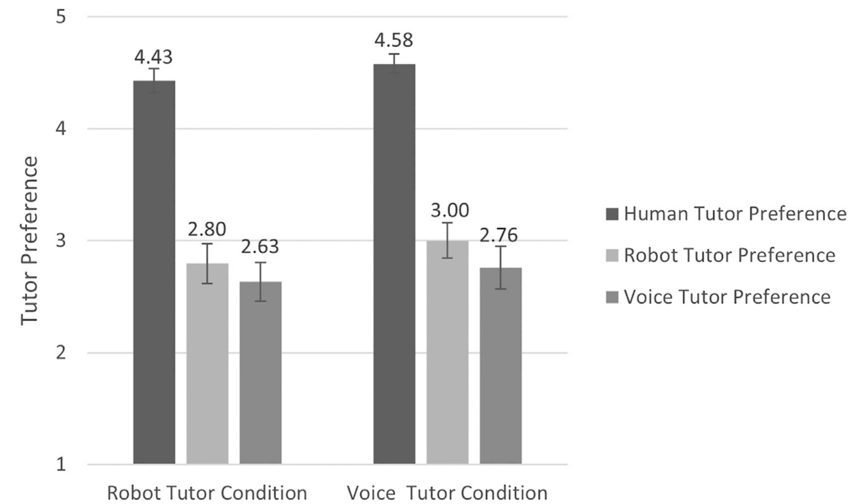


FIGURE 6 | Preference ratings for each tutor after the lesson. *N* = 49 in the robot tutor condition; *N* = 50 in the voice tutor condition. The highest possible score for each test was 5. The error bars indicate the standard errors.

may not be a factor in such non-physical settings (Ligthart and Truong, 2015), hence learning environments with physical materials may yield different results.

Individual Differences and Learning Outcomes

In concert with the previous study concerning in-person lessons (Kanero et al., 2021), we found that negative attitudes toward robots as well as anxiety about learning L2 are related to L2 vocabulary learning with a robot, though the relations were less pronounced. In addition to the measures used previously, we tested the effect of the learner's first impression of the robot tutor. The current study was among the first to test whether 1) the first impressions of the robot affect the learning outcomes, and 2) the impressions of the robot change before and after the interaction. Contrary to our expectation, the first impression ratings did not predict the number of words participants learned from the lesson. Therefore, we found that the NARS, which assessed participants' general attitudes toward robots, was a better predictor of the learning outcomes than the impression of the specific robot tutor.

The inclusion of the impression survey is also relevant for the discussion of the construct validity of the questionnaires used in HRI studies. Many studies used the NARS (Nomura et al., 2006) to measure the attitudes of participants and to predict participants' behaviors (Nomura et al., 2006; Takayama and Pantofaru, 2009; Ivaldi et al., 2017). In both the current study and the previous study (Kanero et al., 2021), although the NARS predicted the number of words participants learned, the correlation was weak to moderate. One possibility was that the difference in generality between the independent variables (i.e., general attitudes toward all robots) and dependent variables (i.e., the number of words learned from a specific robot) led to the relatively weak correlations. Importantly, the impression survey in the current study was a less general measure, but we did not find a correlation between the impression and learning outcomes.

Changes in Attitudes, Impressions, and Preferences From Before to After the Lesson

On average, participants' attitudes toward robots (and voice assistants) became more positive after they interacted with the specific tutor, but the change was not statistically significant. It should be noted that our lesson was very short and the interaction was minimal, and we may expect a greater change when the lesson is longer and more interactive. The NARS was also tested in the previous study (Kanero et al., 2021), and thus we can compare the data of the current study with the data from the in-person lesson. We found that the learner's negative attitudes toward robots did not significantly change before and after the in-person lesson either [t (49) = -1.02, p = 0.31]. As per participants' impressions of the tutor, the first impressions were better for the robot tutor than the voice tutor, but the impressions became comparable between the two conditions after the actual interaction. The results may indicate that, although the impression before the lesson can be affected by embodiment, the short Zoom session was enough for the learners to override the first impressions and assess the agent based on the actual interactive and communicative capabilities. With regard to the learner's preference, we observed a clear preference for a human tutor over both of the machine tutors, and some preference for the robot tutor over the voice tutor. These results also emphasize the importance of choosing different scales depending on what the researcher plans to evaluate.

Limitations and Future Directions

In the current study, embodiment did not facilitate vocabulary learning, and the learner's attitudes toward robots and anxiety about learning L2 consistently predicted learning outcomes. In terms of physical presence, however, we could only compare the current study with the previous study (Kanero et al., 2021) to anecdotally discuss its lack of impact. Therefore, a direct comparison between in-person and virtual lessons should be made before drawing a conclusion. It would also be critical to further test the unique features of robots (e.g., the ability to perform gestures) and to consider other aspects of language such as grammar and speaking (Kanero et al., 2021). Similarly, the lesson scenarios, the demographic characteristics of participants (e.g., education, familiarity with robots) and the morphology of robots (e.g., Pepper, Kismet, Leonardo) might affect learning outcomes. Future research should not only investigate the influence of these factors on learning outcomes, but also analyze the detailed nature of human-robot interaction (e.g., the learner's behaviors during the lesson).

Perhaps most importantly, in the current study, the human-robot interaction was limited to one session lasting only about 15 min. Needless to say, more research is needed to examine whether the physical body of a robot affects learning outcomes in other settings such as a lesson on another subject, or in a longer and more interactive lesson. The effects of embodiment may be more pronounced when multiple lessons are provided over a longer period of time. Further, some researchers suggest that robot tutors may reduce the L2 anxiety of child learners in the long run (Alemi et al., 2015), and thus future research may focus on the long-term effects of robot language lessons on the anxiety levels of children and adults. Another recent study also found that children between 5 and 6 years old do not interact with voice assistants as much as they interact with humans (Aeschlimann et al., 2020). To our knowledge, no child study has compared robots and voice assistants. Overall, developmental research should adopt an experimental design similar to our study and examine whether the current findings can be replicated with a younger population.

Our data in the voice tutor condition also provide insights into the effectiveness of voice assistants such as Amazon Alexa and Apple Siri. Research with children suggests that voice assistants are perceived as a source of information that can answer questions about a wide range of subjects, including language such as definitions, spellings, and translations (Lovato et al., 2019). Our results show that adults can learn a second language from voice assistants as well, at least to the same extent they do with social robots. It should also be noted that one reason why we did not find a link between negative attitudes toward voice assistants and learning outcomes might be that we adapted a questionnaire about robots, simply by changing the word "robot" to "voice assistant." While this manipulation made the two conditions as comparable as possible, the validity of the voice assistant questionnaire should be carefully considered. Future research may use our findings as a base to explore how and for whom voice tutors are beneficial.

Finally, we should also point out that the current study was conducted amid the COVID-19 pandemic. We believe that our findings are generalizable, and if anything, the pandemic might have provided a better setting to evaluate the impact of (dis)embodiment. Online education has become abundant, and people may be less hesitant to engage in virtual interactions, hence the difference between in-person and online interactions should be less driven by the unfamiliarity of online interactions in the current climate. Nevertheless, more studies should be conducted to critically assess the generalizability of the findings.

Conclusion

This study was the first to empirically investigate the influence of the robot's physical embodiment on second language learning. The study presents an example of embodiment not affecting the learning outcomes although the results should be interpreted cautiously until the results are replicated for different language learning tasks and using various scenarios and interaction designs. Evaluating the influences of individual differences in robot-led Zoom lessons, we also found that the learner's general attitudes toward robots predict learning outcomes. Our findings provide some hope for the difficult situation during the COVID-19 pandemic because participants successfully learned vocabulary in a short Zoom lesson. The current results also encourage more researchers to be engaged in studying the influence of the user's individual differences in human-robot interaction and policymakers and educators to carefully consider how social robots and other technological devices should be incorporated in educational settings.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study will be available from the corresponding author on reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Sabancı University. The participants completed an online consent form to participate in this study.

AUTHOR CONTRIBUTIONS

JK conceived the study in consultation with ET and CO. JK and ET were in charge of collecting the data. ET and JK analyzed the data in consultation with CO, TG, and AK. JK, ET, and CO drafted the manuscript, and all authors critically edited it. All authors contributed to the project and approved the final submitted version of the manuscript.

FUNDING

The publication fee of the manuscript was paid from the institutional research funds of Sabancı University and Koç University.

ACKNOWLEDGMENTS

We would like to thank the members of Sabancı University Mind, Language, & Technology Lab – Serap Özlü for creating study materials and conducting the experiment, and İrem Gözelekli, Ceren Boynuk, and Ayşenaz Akbay for conducting the experiment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.679893/full#supplementary-material

Video 1.MP4 | An excerpt from the robot tutor lesson.

Video 2.MP4 | An excerpt from the voice tutor lesson.

REFERENCES

Admoni, H., and Scassellati, B. (2017). Social Eye Gaze in Human-Robot Interaction: A Review. *J. Hum.-Robot Interact.* 6, 25. doi:10.5898/JHRI.6.1.Admoni

Aeschlimann, S., Bleiker, M., Wechner, M., and Gampe, A. (2020). Communicative and Social Consequences of Interactions with Voice Assistants. *Comput. Hum. Behav.* 112, 106466. doi:10.1016/j.chb.2020.106466

Alemi, M., Meghdari, A., and Ghazisaedy, M. (2015). The Impact of Social Robotics on L2 Learners' Anxiety and Attitude in English Vocabulary Acquisition. *Int. J. Soc. Robotics* 7, 523–535. doi:10.1007/s12369-015-0286-y

Aydın, S., Harputlu, L., Güzel, S., Çelik, S. S., Uştuk, Ö., and Genç, D. (2016). A Turkish Version of Foreign Language Anxiety Scale: Reliability and Validity. *Proced. - Soc. Behav. Sci.* 232, 250–256. doi:10.1016/j.sbspro.2016.10.011

Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., and Ruppert, J. A. (2003). The Psychology of Learning and Motivation: Advances in Research and Theory, *Elsevier Science*, 43–92. doi:10.1016/s0079-7421(03)01011-9 Social Embodiment.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social Robots for Education: A Review. *Science Rob.*, 3. doi:10.1126/scirobotics.aat5954

Bartneck, C., and Forlizzi, J. (2004). "A Design-Centred Framework for Social Human-Robot Interaction," in *RO-MAN 2004: 13th IEEE International Workshop on Robot and Human Interactive Communication*. (Kurashiki, Japan: Institute of Electrical and Electronics Engineers), 591–594. doi:10.1109/ROMAN.2004.1374827

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Statistical Software* 67, 1–48. doi:10.18637/jss.v067.i01

Bialystok, E., Craik, F. I. M., Klein, R., and Viswanathan, M. (2004). Bilingualism, Aging, and Cognitive Control: Evidence from the Simon Task. *Psychol. Aging* 19, 290–303. doi:10.1037/0882-7974.19.2.290

Chevalier, P., Martin, J.-C., Isableu, B., Bazile, C., and Tapus, A. (2017). Impact of Sensory Preferences of Individuals with Autism on the Recognition of Emotions Expressed by Two Robots, an Avatar, and a Human. *Auton. Robot* 41, 613–635. doi:10.1007/s10514-016-9575-z

Demir, B., and Kumkale, G. T. (2013). Individual Differences in Willingness to Become an Organ Donor: A Decision Tree Approach to Reasoned Action. *Personal. Individual Differences* 55, 63–69. doi:10.1016/j.paid.2013.02.002

Deng, E., Mutlu, B., and Mataric, M. J. (2019). Embodiment in Socially Interactive Robots. *FNT in Robotics* 7 (4), 251–356. doi:10.1561/2300000056

Dunn, L. M., and Dunn, D. M. (2007). *Peabody Picture Vocabulary Test(PPVT)*. Fourth edition (Minneapolis, MN: NCS Pearson).

Gunaydin, G., Selcuk, E., and Zayas, V. (2017). Impressions Based on a Portrait Predict, 1-Month Later, Impressions Following a Live Interaction. *Soc. Psychol. Personal. Sci.* 8, 36–44. doi:10.1177/1948550616662123

Han, J., Jo, M., Jones, V., and Jo, J. H. (2008). Comparative Study on the Educational Use of home Robots for Children. *J. Inf. Process. Syst.* 4, 159–168. doi:10.3745/JIPS.2008.4.4.159

Hoff, E. (2013). Interpreting the Early Language Trajectories of Children from Low-SES and Language Minority Homes: Implications for Closing Achievement Gaps. *Dev. Psychol.* 49, 4–14. doi:10.1037/a0027238

Horwitz, E. K., Horwitz, M. B., and Cope, J. (1986). Foreign Language Classroom Anxiety. *Mod. Lang. J.* 70, 125–132. doi:10.2307/32731710.1111/j.1540-4781.1986.tb05256.x

Ivaldi, S., Lefort, S., Peters, J., Chetouani, M., Provasi, J., and Zibetti, E. (2017). Towards Engagement Models that Consider Individual Factors in HRI: On the Relation of Extroversion and Negative Attitude towards Robots to Gaze and Speech during a Human-Robot Assembly Task. *Int. J. Soc. Robotics* 9, 63–86. doi:10.1007/s12369-016-0357-8

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and Towards Logit Mixed Models. *J. Mem. Lang.* 59, 434–446. doi:10.1016/j.jml.2007.11.007

Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., and Göksun, T. (2018). Social Robots for Early Language Learning: Current Evidence and Future Directions. *Child. Dev. Perspect.* 12, 146–151. doi:10.1111/cdep.12277

Kanero, J., Oranç, C., Koşkulu, S., Kumkale, G. T., Göksun, T., and Küntay, A. C. (2021). Are Tutor Robots for Everyone? the Influence of Attitudes, Anxiety, and Personality on Robot-Led Language Learning. *Int. J. Soc. Robotics.* doi:10.1007/s12369-021-00789-3

Kennedy, J., Baxter, P., and Belpaeme, T. (2015). Comparing Robot Embodiments in a Guided Discovery Learning Interaction with Children. *Int. J. Soc. Robotics* 7, 293–308. doi:10.1007/s12369-014-0277-4

Konishi, H., Kanero, J., Freeman, M. R., Golinkoff, R. M., and Hirsh-Pasek, K. (2014). Six Principles of Language Development: Implications for Second Language Learners. *Dev. Neuropsychol.* 39, 404–420. doi:10.1080/87565641.2014.931961

Kovács, A. M., and Mehler, J. (2009). Cognitive Gains in 7-Month-Old Bilingual Infants. *Proc. Natl. Acad. Sci.* 106, 6556–6560. doi:10.1073/pnas.0811323106

Köse, H., Uluer, P., Akalin, N., Yorganci, R., Özkul, A., and Ince, G. (2015). The Effect of Embodiment in Sign Language Tutoring with Assistive Humanoid Robots. *Int J. Soc. Rob.* 7, 537–548. doi:10.1007/s12369-015-0311-1

Lovato, S. B., Piper, A. M., and Wartella, E. A. (2019). "Hey Google, Do Unicorns Exist?: Conversational Agents as a Path to Answers to Children's Questions," in *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (Boise, ID USA: ACM), 301–313. doi:10.1145/3311927.3323150

Leyzberg, D., Spaulding, S., Toneva, M., and Scassellati, B. (2012). The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. *Proc. Annu. Meet. Cogn. Sci. Soc.* 6.

Li, J. (2015). The Benefit of Being Physically Present: A Survey of Experimental Works Comparing Copresent Robots, Telepresent Robots and Virtual Agents. *Int. J. Human-Computer Stud.* 77, 23–37. doi:10.1016/j.ijhcs.2015.01.001

Ligthart, M., and Truong, K. P. (2015). "Selecting the Right Robot: Influence of User Attitude, Robot Sociability and Embodiment on User Preferences," in *24th IEEE International Symposium on Robot and Human Interactive Communication*. (Kobe, Japan: ROMAN: IEEE), 682–687. doi:10.1109/ROMAN.2015.7333598

Lytle, S. R., and Kuhl, P. K. (2017). "Social Interaction and Language Acquisition: Toward a Neurobiological View," in *The Handbook of Psycholinguistics*. Editors E. M. Fernández and H. S. Cairns (Hoboken, NJ: Wiley Blackwell), 615–634. doi:10.1002/9781118829516.ch27

Macedonia, M., Groher, I., and Roithmayr, F. (2014). Intelligent Virtual Agents as Language Trainers Facilitate Multilingualism. *Front. Psychol.* 5. doi:10.3389/fpsyg.2014.00295

Milligan, K., Astington, J. W., and Dack, L. A. (2007). Language and Theory of Mind: Meta-Analysis of the Relation Between Language Ability and False-belief Understanding. *Child Develop.* 78, 622–646. doi:10.1111/j.1467-8624.2007.01018.x

Nomura, T., Kanda, T., and Suzuki, T. (2006). Experimental Investigation into Influence of Negative Attitudes toward Robots on Human-Robot Interaction. *AI Soc.* 20, 138–150. doi:10.1007/s00146-005-0012-7

Paolo, A. D., and Tansel, A. (2015). Returns to Foreign Language Skills in a Developing Country: The Case of Turkey. *J. Dev. Stud.* 51, 407–421. doi:10.1080/00220388.2015.1019482

Pfeifer, R., and Scheier, C. (1999). *Understanding Intelligence*. MIT Press.

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. Available at: http://www.r-project.org/.

Rosenthal-von der Pütten, A. M., Straßmann, C., and Krämer, N. C. (2016). "Robots or Agents - Neither Helps You More or Less during Second Language Acquisition," in *Intelligent Virtual Agents Lecture Notes in Computer Science.*. Editors D. Traum, W. Swartout, P. Khooshabeh, S. Kopp, S. Scherer, and A. Leuski (Cham: Springer International Publishing), 256–268. doi:10.1007/978-3-319-47665-0_23

Takayama, L., and Pantofaru, C. (2009). Influences on Proxemic Behaviors in Human-Robot Interaction. *IEEE/RSJ Int. Conf. Intell. Robots Syst.* 5495–5502. doi:10.1109/IROS.2009.5354145

University of Cambridge Local Examinations Syndicate (2001). *Quick Placement Test*. Oxford, UK: Oxford University Press.

Verga, L., and Kotz, S. A. (2013). How Relevant Is Social Interaction in Second Language Learning?. *Front. Hum. Neurosci.* 7. doi:10.3389/fnhum.2013.00550

Willis, J., and Todorov, A. (2006). First Impressions. *Psychol. Sci.* 17, 592–598. doi:10.1111/j.1467-9280.2006.01750.x

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

ORIGINAL RESEARCH
published: 06 July 2021
doi: 10.3389/frobt.2021.692811

Becoming Team Members: Identifying Interaction Patterns of Mutual Adaptation for Human-Robot Co-Learning

Emma M. van Zoelen[1,2]*, *Karel van den Bosch*[2] *and Mark Neerincx*[1,2]

[1]Interactive Intelligence, Intelligent Systems Department, Delft University of Technology, Delft, Netherlands, [2]Human-Machine Teaming, Netherlands Organization for Applied Scientific Research (TNO), Soesterberg, Netherlands

Becoming a well-functioning team requires continuous collaborative learning by all team members. This is called *co-learning*, conceptualized in this paper as comprising two alternating iterative stages: partners adapting their behavior to the task and to each other (co-adaptation), and partners sustaining successful behavior through communication. This paper focuses on the first stage in human-robot teams, aiming at a method for the identification of recurring behaviors that indicate co-learning. Studying this requires a task context that allows for behavioral adaptation to emerge from the interactions between human and robot. We address the requirements for conducting research into co-adaptation by a human-robot team, and designed a simplified computer simulation of an urban search and rescue task accordingly. A human participant and a virtual robot were instructed to discover how to collaboratively free victims from the rubbles of an earthquake. The virtual robot was designed to be able to real-time learn which actions best contributed to good team performance. The interactions between human participants and robots were recorded. The observations revealed patterns of interaction used by human and robot in order to adapt their behavior to the task and to one another. Results therefore show that our task environment enables us to study co-learning, and suggest that more participant adaptation improved robot learning and thus team level learning. The identified interaction patterns can emerge in similar task contexts, forming a first description and analysis method for co-learning. Moreover, the identification of interaction patterns support awareness among team members, providing the foundation for human-robot communication about the co-adaptation (i.e., the second stage of co-learning). Future research will focus on these human-robot communication processes for co-learning.

Keywords: human-robot collaboration, human-robot team, co-learning, co-adaptation, interaction patterns, emergent interactions

OPEN ACCESS

Edited by:
*Tony Belpaeme,
Ghent University, Belgium*

Reviewed by:
*Christopher Crick,
Oklahoma State University,
United States
Javier de Lope,
Polytechnic University of Madrid,
Spain*

***Correspondence:**
*Emma M. van Zoelen
e.m.vanzoelen@tudelft.nl*

Specialty section:
*This article was submitted to
Human-Robot Interaction,
a section of the journal
Frontiers in Robotics and AI*

Received: *09 April 2021*
Accepted: *16 June 2021*
Published: *06 July 2021*

Citation:
van Zoelen EM, van den Bosch K and Neerincx M (2021) Becoming Team Members: Identifying Interaction Patterns of Mutual Adaptation for Human-Robot Co-Learning. Front. Robot. AI 8:692811. doi: 10.3389/frobt.2021.692811

INTRODUCTION

When people collaborate in teams, it is of key importance that all team members get to know each other, explore how they can best work together, and eventually adapt to each other and learn to make their collaboration as fluent as possible. While humans do this naturally (Burke et al., 2006), it is not self-evident for robots that are intended to function as team partners in human-robot collaborations.

It is well known that robotic team partners should be transparent, predictable, and explainable, but it is often overlooked that human team partners *become* predictable and explainable through a process of exploration and mutual learning.

We call the above mentioned process co-learning (Bosch et al., 2019). While existing work on human-robot collaboration and mutual adaptivity often focuses on short-term single interactions, we believe it is necessary to also look at repeated interactions to study co-learning as a mechanism for building fluent human-robot collaborations. We conceptualize co-learning as comprising two alternating iterative stages. In the *first* stage, partners observe each other and adapt their behavior to the other, leading to successful emergent team behaviors. Such adaptation can be done deliberately but often occurs implicitly and unconsciously. In the *second* stage, partners communicate about their adaptations and give each other feedback, thereby giving meaning to and becoming aware of the learned behavior. Especially this second stage of creating awareness of what has been learned helps to sustain the behavioral adaptations over time and across contexts.

We regard co-learning to be vital for creating successful human-robot collaborations. However, the term "co-learning" is relatively new in human-robot interaction literature, and it is not yet precisely defined what human-robot co-learning looks like in practice and how it should be studied. Emergent behavior can only be investigated through empirical studies; to investigate human-robot co-learning it is therefore necessary that *both* partners can learn in real-time while collaborating with each other. In this study, we therefore chose to empirically study co-learning with a human participant and a Reinforcement Learning (RL) virtual robot. For the investigation of emerging co-adaptive behaviors, we distinguish four main research questions:

1. How to identify and classify recurring behaviors that indicate co-learning in a human-robot team?
2. Which recurring sequences of these behaviors (*co-learning patterns*) can be identified, such that they can be used by the team partners to communicate about their adaptations?
3. How does the robot's learning, emerging from the interactions, affect the human's behavior and learning?
4. How does the human's learning, emerging from the interactions, affect the robot's behavior and Reinforcement Learning?

The literature on human-robot interaction and learning contains a large body of research on personalized robot tutors, in which a robot tutor learns to personalize its interactions to support the learning process of a human student, focused on classroom or training related contexts [a few examples are (Baxter et al., 2017; Gao, Barendregt, and Castellano 2017; Belpaeme et al., 2018; Vignolo et al., 2019)]. Formal training is important to initiate learning and to steer development in the right direction. However, it is important to realize that learning continues after training has been completed. Every new experience in the task provides an opportunity for human-robot teams to learn from their collaboration. An important aspect of co-learning in actual task contexts is developing and refining (shared) mental models about team members and about the task at hand, to increase mutual understanding of the best way to perform the task (Klein et al., 2004). Therefore, we specifically attempt to answer the above mentioned questions in a task context where learning happens during task execution.

In this paper, we present a behavioral study of how a human and a virtual robot, which uses a Reinforcement Learning algorithm to adapt and optimize its actions, adapt their behavior to collaboratively solve a task. We are interested in how the behavior of the human ánd the behavior of the robot changes as a result of this process, making our study fit within a new area of research in which both human and machine behavior are assessed [with their mutual dependencies; cf. (Rahwan et al., 2019)]. We first provide an in-depth elaboration of the concept "co-learning" within the context of human-robot collaboration, resulting in a definition of co-learning and the related concepts of co-adaptation and co-evolution. Based on this, we identify the requirements for empirical research into co-adaptation and co-learning, and present the design of an environment for studying co-learning. This environment has been built and used to conduct an empirical study into human-robot co-learning. From the analysis of the observed human-robot interactions, a list of patterns of adaptive interactions, and the switching between these patterns over time, were identified. These patterns can emerge in similar task contexts, forming a first description method for co-learning analyses. Moreover, it supports the creation of awareness, providing the foundation (concepts) for human-robot communication about the co-adaptation (i.e., the second stage of co-learning).

CO-LEARNING: BACKGROUND AND DEFINITION

Collaborative learning is a widely studied mechanism in human-only contexts, and it was Dillenbourg (Dillenbourg et al., 1996) who suggested that collaborative learning can also take place between humans and computers. Collaborative learning in the context of Dillenbourg's work means that learning (the acquisition of new knowledge, skills, behavior, etc.) results from collaborative activities between team partners. If we look at human-robot interaction literature, several terms are used that describe a similar process in which two parties or systems change their behavior and/or mental states concurrently while interacting with each other. Co-adaptation (Xu et al., 2012; Chauncey et al., 2016; Nikolaidis et al., 2017a) and co-learning (Bosch et al., 2019) are two of them, but we also encounter co-evolution (Döppner et al., 2019), in which "co" stands for collaborative, also meaning "mutual."

Co-adaptation and co-learning are often used interchangeably, making it difficult to understand what they stand for. There are several vision papers explaining the importance of both co-adaptation [e.g. (Xu et al., 2012; Ansari, Erol, and Sihn 2018)] as well as co-learning [e.g. (Bosch et al., 2019; Holstein, Aleven, and Rummel 2020; Wenskovitch and North 2020)], but a clear distinction between the two, or a definition specifically for co-learning, is missing from these papers. When looking at the experimental literature, however, it seems that there are subtle

TABLE 1 | The concepts co-adaptation, co-learning and co-evolution defined in terms of timespan in which they occur, persistence and intention.

	Co-adaptation	Co-learning	Co-evolution
Timespan	Short (seconds—hours)	Medium (hours—weeks)	Long (weeks—years)
Persistence	Developed behavior/mental state does not necessarily persist over time, and probably not at all across contexts	Developed behavior/mental state persists over time and possibly across contexts	Developed behavior/mental state might persist for a while but possibly continues to evolve, similar to the development of habituation
Intention	Changes and developments happen as a consequence of interactions and an implicit or explicit drive to improve performance or experience	Explicitly goal-driven: Attempts to improve performance or experience; learning is an explicit goal	Changes and developments happen as a consequence of interactions and possibly an implicit drive to improve performance or experience

differences. Experimental studies on co-adaptation often focus on making the agent or robot adaptive to the human, using different kinds of information about the human [e.g. (Buschmeier and Kopp 2013; Ehrlich and Cheng 2018; Sordoni et al., 2015; Yamada and Yamaguchi 2002)]. Some studies have investigated how a human adapts in situations in which they collaborate with an intelligent agent or robot. These studies mostly focus on the performance of the human and their resulting subtle behavior change in short experiments [e.g. (Nikolaidis et al., 2017b; Mohammad and Nishida., 2008; Nikolaidis et al., 2017a)]. The studies that use "co-learning" tend to take a more symmetrical approach by looking at agent or robot learning as well as human learning, and pay more attention to the learning process and changing strategies of the human as well, often looking at many repetitions of a task (Ramakrishnan, Zhang, and Shah 2017; C.-S. Lee et al., 2020; C. Lee et al., 2018; Shafti et al., 2020). Studies on co-evolution, on the other hand, monitor a long-term real-world application in which behavior of the human as well as the robot subtly changes over time (Döppner, Derckx, and Schoder 2019).

Following these differences, we propose to distinguish the terms using three dimensions, namely 1) the time over which the development takes place, 2) the persistence of the resulting behavior/mental state over time and across contexts, and 3) the intention of the development. **Table 1** shows the proposed definitions in detail. Within our research, we focus on co-learning as defined here.

In a human-robot co-learning process, a human and a robot collaborate on a given task. In order to do well on the task, they need to learn all kinds of implicit and explicit knowledge related to both the task itself as well as the collaboration and interaction between them. Related to the task they can, for example, learn the technical details of how the task should be executed. Related to the collaboration, they can, for example, learn social collaboration skills. Related to both, they can learn about their own role and the role of the other in the task and the consequences of their own and their partner's actions and mental state on the task (how to collaborate in context of the task). Ultimately, learning this should help them to together perform well on the task, to build understanding of each other in context of the task and to calibrate the trust that the human and the robot have in each other. We focus our work on this last type of combined task and collaboration learning.

We further define co-learning to be comprised of two stages that follow each other in continuous iterative cycles, namely 1) co-adaptation, and 2) a communication process. Part (a) is therefore a process in which team members (sometimes unconsciously) adapt to each other and the task, thereby changing and developing their behavior as a consequence of interactions and an implicit or explicit drive to improve performance or experience (see again **Table 1**). Part (b) is a process in which these implicitly developed behaviors are shared and discussed through direct communications or interactions between team members, thereby making the team members aware of the implicit adaptations. This combination ensures that learned strategies are grounded in the context and task and can be strategically used in new contexts.

RESEARCH CHALLENGES

Many research challenges follow from the conceptualization of co-learning, due to the fact that both human and robot are non-static. They are both constantly developing, changing and adapting, and they influence each other in the process. This means that it is not possible to study only one of the team partners; it is necessary to take a symmetrical approach, where both human and robot are studied through the interactions between them. Moreover, co-learning in dynamic tasks is a continuous process in which new task situations that appear dynamically require new learning over and over again. Therefore, focusing on one specific interaction, or on team performance as end result, does not offer a complete picture. We need to study all interactions that contribute to this process. These specific dynamic properties need to be taken into account in the design of experiments, as well as in the analysis and discussion of results. Following from this, and to provide a broader view on the specific study that we present in this paper, we have defined three research directions that need to be addressed in the study of human-robot co-learning:

1. **Research into enabling and assessing co-learning:** to understand the dynamics of co-learning, we need to investigate what kind of behaviors and interactions drive co-adaptation and co-learning, and how learning processes of human and robot team members influence each other.
2. **Research into interaction patterns that make team partners explicitly aware of learned behavior, such that behavior can be sustained over time and context:** in order to create sustainable team behavior, human and robot need to communicate about learned behavior to ensure that they are aware of useful learned behavior. It is important to investigate what kind of communication interaction patterns enable this specific type of communication.

3. **Research into a dynamic team mental model that takes into account naturally occurring changes in interaction patterns, and how such a model can support the robot in its learning process:** as humans, we are able to anticipate on the fact that our team members learn and change. It is important to investigate how a dynamic team model can enable robots to also anticipate the fact that their human team member is continuously changing.

The study presented in this paper focuses on the first research direction; the research questions presented in the introduction have been derived from it. More specifically, in the experiment that we describe in the following sections, we have chosen to focus on the first stage of the co-learning process: co-adaptation as a precursor for co-learning. We do not yet address questions concerning communication about learned behaviors (research direction 2), but focus on the implicit behavioral adaptations that occur within a relatively short time span. It is expected that results of the present study will provide pointers for how to investigate the issues associated with research challenges two and three above.

RESEARCH ENVIRONMENT: DESIGNING TASK, AGENT, CONTEXT

Context

To study co-learning in human-robot teams, a suitable task context needs to be designed. We identify the following requirements for such a task context in which we can study co-adaptation according to the definition in **Table 1**:

1. It should be possible for the team to improve its performance by making effective use of the capabilities of the individual team members [as this is necessary to make it a team task (Johnson et al., 2014)];
2. There should be possibilities for implicit adaptation and learning for both human and robot team members;
3. It should accommodate different emergent collaborative strategies for solving the task;
4. For this first study, the task and team work should be simple enough for a Reinforcement Learning agent to learn new behavior in a short number of rounds, such that we can study co-adaptation in relatively short experimental sessions;
5. To ensure societal relevance of this research, the task should be based on a real-life domain in which there is a need for autonomous robots that function as team partners.

As a general context for defining a task, we chose Urban Search and Rescue (USAR). A lot of research on human-agent teaming is done in USAR-related tasks (Lematta et al., 2019), because the safety-critical nature makes the application of human-robot collaboration very useful; there are ongoing initiatives which aim to use robots in real USAR teams (requirement 5). Moreover, it is a dynamic task context with many possible subtasks and possibilities for the introduction of threats, safety risks and changing information.

We developed an earthquake scenario for our human-robot USAR team, with the team's task to remove rubble and debris from a victim. To get a better understanding of the task, and of the knowledge and capabilities it requires from partners, we created a storyboard (**Figure 1**). The storyboard shows a possible scenario in which the robot picks up a large rock to clear it away, not realizing that the action may lead to a small rock falling on the head of the victim (**Figure 1C**). When the human notices this, this provides an opportunity to jump in, and to prevent the rock from crushing the victim (**Figure 1D**). This event facilitates the human learning that the robot apparently does not understand the risks of falling rocks. It provides the robot with the opportunity to learn that it made a mistake. The event furthermore provides an opportunity for team members to communicate about the event, their actions, and to plan how they will manage such situations together in the future. This storyboard illustrates that using the unique capabilities of both the human (insight into strategic choices) and robot (physical strength) can be exploited to achieve better task performance (requirement 1). The task of removing rubble from a victim allows for a great diversity in task planning and execution, and for the development of individual strategies (requirement 2 and 3), as the different debris can differ in shape, size and location, while enabling simple basic actions to create strategies for solving the task (requirement 4).

Task Implementation

We developed a digital task simulation of the described USAR context using Python and the MATRX package (MATRX Software, 2020). MATRX is a package for rapid prototyping of human-agent team environments, which supports easy generation of an environment, object and agents. **Figure 2** shows a screenshot of the simulation. The scene involves three characters: a victim buried underneath a pile of rocks (shown in the middle), an explorer (avatar on the left, played by a human participant) and a Reinforcement Learning robot agent (avatar on the right). The goal of the task is to free the victim by clearing away all rocks that are in front of the victim, as well as to create a pathway to the victim from either the left or the right side. In order to score well, this must happen as quickly as possible, and no additional rocks (or as little as possible) should fall on top of the victim, as that will cause extra harm. Both the human and the virtual robot each have a set of actions they can perform, such as picking up rocks and dropping them somewhere else. However, the extent to which they can perform actions differs: the robot can pick up large and small objects, and break large objects into pieces. Humans can only pick up small objects. Humans however have a better insight in certain aspects of the task that dictate which actions are useful to do, such as how rocks will fall when other rocks are removed or replaced. This insight stems from the fact that humans have "common sense," which helps us understand the probable consequences of actions. In order to complete the task successfully participants must collaborate with the virtual robot (requirement 1), while managing their actions in such a way that the robot does not accidentally drop rocks on the victim's head. Since it is not clear at the start what the best strategy would be to solve the task quickly, both partners need to learn and adapt as they go (requirement 2). The levels are designed such that there are different possible ways to solve the task (requirement 3), and it is a discrete environment build on a simple state machine, making it possible to design a Reinforcement Learning agent that can process the environment (requirement 4).

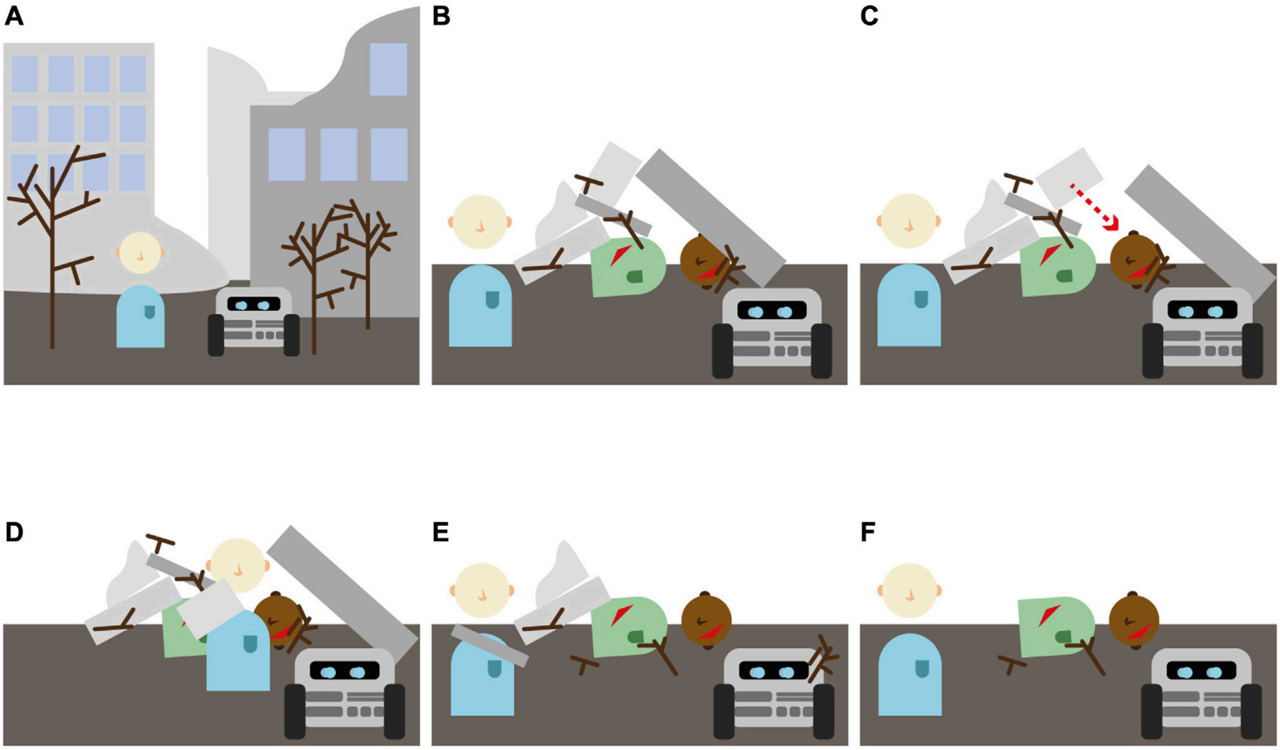


FIGURE 1 | A storyboard describing how a human-robot team might free an earthquake victim from underneath a pile of rocks. In this storyboard, we can see how the robot picks up a large rock, unaware that this will cause another rock to fall on the head of the victim **(panel C).** The human notices the issue, and steps in to prevent the rock from falling **(panel D).** This event can help the robot learn about the task, and that it apparently made a mistake. The human can learn about the capabilities of the robot, namely that it didn't understand how the rocks would fall and that it would cause harm.

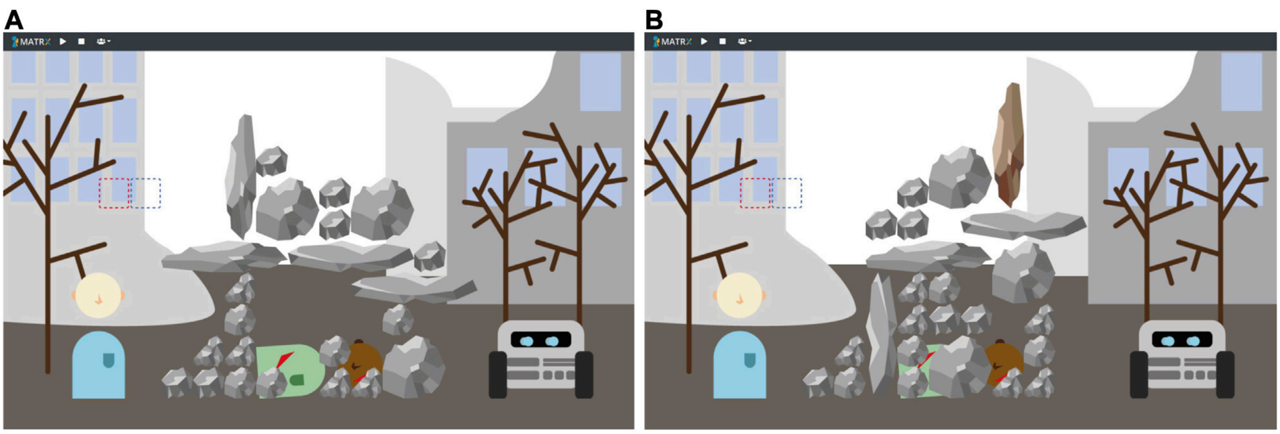


FIGURE 2 | The USAR task environment programmed in MATRX. It shows a victim underneath a pile of rocks, and a human and a robot representing the team members. The dashed red square (above the human's head) represents the hand of the human that can be moved to pick up rocks. The dashed blue square represents the hand of the robot. Scene **(A)** was used as level 1 in the experiment, while scene **(B)** was used as level 2.

Learning Agent

To be able to empirically study how a human and a robot co-adapt while collaborating on a task, the robot should be able to try out and evaluate different actions, to be able to choose the policy that best fits the goals of the team given the adaptations done by the human team member. We chose to use Reinforcement Learning to enable the agent to learn, for three main reasons:

1. The robot had to be able to learn real-time, on the basis of rewards: as the behavior of the human team partner is adaptive

TABLE 2 | The task conditions specified for each Phase Variable used in the state space of the Reinforcement Learning algorithm.

Phase	Description
Phase 1	The starting phase: Describes the state of the task environment when no rocks have been moved
Phase 2	The heights of all piles of rocks added up is now at least 10 rocks lower than in phase 1
Phase 3	Phase 2 has been reached, and the heights of all piles of rocks added up is now at least 20 rocks lower than in phase 1
Phase 4	Phase 2 and 3 have been reached, and either there are no more rocks directly on top of the victim, OR one of the sides of the task field is cleared from rocks, meaning there is an access route to the victim from either the left or right side
Goal phase	Phase 2, 3 and 4 have been reached, and there are no more rocks directly on top of the victim, AND one of the sides of the task field is cleared from rocks, meaning there is a free route from either the left or right side to the victim. The task terminates when this phase is reached

and unpredictable, we cannot determine optimal behavior before the start of the task. This means that the best way to find the optimal strategy for solving the task collaboratively would be to get feedback or rewards on performance.
2. The described task can be solved by a sequence of actions that manipulate the state of the world with each action, therefore, the learning algorithm had to be able to learn a sequence of actions given different sequential task situations.
3. The described task is a human-robot scenario, but it can also conceptually be seen as a multi-agent scenario as both the robot and the human are autonomous and learning agents within the collaboration. Reinforcement Learning is an often used and widely studied mechanism in multi-agent scenarios, for reasons related to reasons one and two above as well [see e.g. (Kapetanakis and Kudenko 2002; Foerster et al., 2016)].

Reinforcement Learning has been designed for learning sequences of actions in tasks that can be modeled as Markov Decision Processes (van Otterlo and Wiering 2012), in which the transitions between states are unknown. In contexts where agents collaborate and learn with a human, these transitions are unknown since it is unknown what the human will do; this is also the case in our task. While such a human-agent collaborative context poses many challenges (e.g. large state spaces, long convergence times and random behavior in the beginning) (Dulac-Arnold, Mankowitz, and Hester 2019), earlier work has shown that RL can be used successfully for learning behavior in real time when interacting with a human, provided that the learning problem is simple enough (Weber et al., 2018). Since we used RL mostly as a tool to ensure that the agent could adapt over time, and not as a goal in itself, we created a RL mechanism that is much simpler than the current state-of-the-art, but that would provide the basic learning that is sufficient for our research goals. We simplified the task by modeling it as a semi-Markov Decision Process (Sutton, Precup, and Singh 1999), which means that the task is divided into several "phases," which serve as the states in the RL algorithm. Normally states last one timestep, whereas in a semi-Markov Decision Process, these phases can last variable amounts of time. Our state definition describes the state of the environment based on the amount of rocks present in the area around the victim. The state space is defined by S = (Phase 1, Phase 2, Phase 3, Phase 4, Goal Phase). **Table 2** describes the details of the individual states. We chose to not explicitly represent learning about collaboration in the learning agent, since we wanted to focus on implicit behavioral adaptations (as explained in *Research Challenges*). We combined this with a system inspired by the Options Framework (Stolle and Precup 2002) and a basic greedy Q-learning algorithm. In the Options Framework, agents use RL to learn a meta-policy as well as several "sub-policies." These sub-policies can also be seen as macro-actions; they are combinations of atomic actions that are used together to solve parts of the task. Usually, these macro-actions are learned in parallel with the meta-policy, but sometimes they are pretrained, such as for example in (Illanes et al., 2019). To further simplify the learning problem, we chose to predefine three rule-based macro-actions; the agent could choose from these macro-actions in each phase of the task (a description of each macro-action is given in **Figure 3**, **Figure 4**, and **Figure 5**). The rewards for the RL algorithm are based on two factors: 1) the time it took the team to move to the next phase, and 2) the amount of additional harm done to the victim. The agent would receive this reward when transitioning into a new phase, or when the task terminates due to becoming unsolvable or due to a timeout. The height of the rewards was made such that the total reward given was always negative. With initial Q-values of 0, this ensured that in the first three runs of the experiment, the agent would try out all three macro-strategies in order, to enforce initial exploration. A visual overview of the learning problem is provided in **Figure 6**, after we have explained more details about the experimental method.

Claims: Expected Observations

We expect to observe several behaviors within this task environment, given that it was designed to study co-learning behavior. We have formulated these expected observations as the following claims:

- Different participants develop different ways of performing the task;
- The agent learns different sequences of macro-actions for different participants;
- Different teams converge to different ways of performing the task;
- The agent converges to a specific sequence of macro-actions for most participants;
- The human converges to a specific strategy within the experiment.

In the Discussion (*Discussion*), we use the results of our experiment to critically evaluate whether we have been able to study co-adaptation as a precursor for co-learning with our methods by verifying to what extent these claims hold.

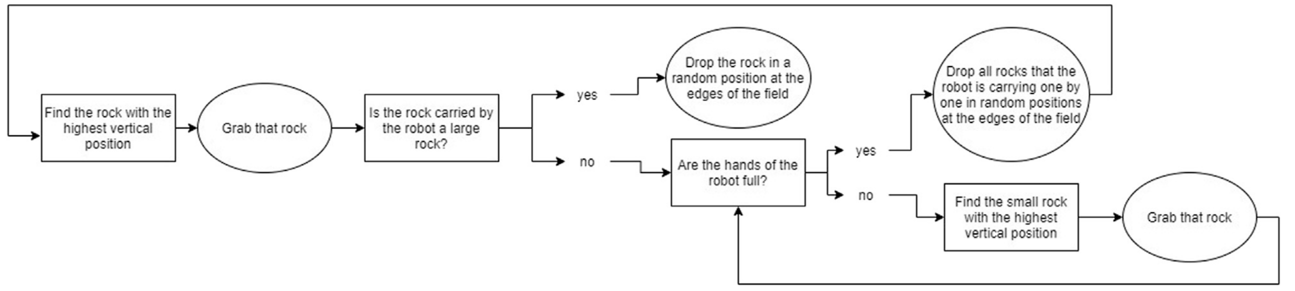


FIGURE 3 | A flowchart showing the rule-based decision making the agent would go through when using Macro-Action 1.

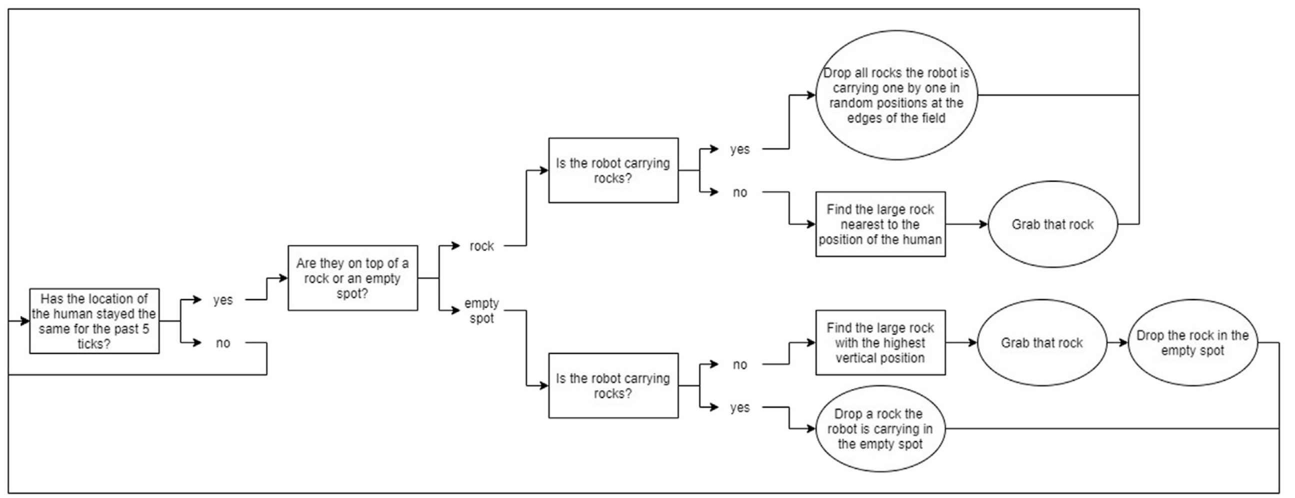


FIGURE 4 | A flowchart showing the rule-based decision making the agent would go through when using Macro-Action 2.

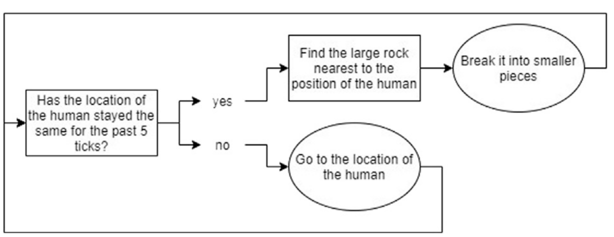


FIGURE 5 | A flowchart showing the rule-based decision making the agent would go through when using Macro-Action 3.

METHOD OF STUDY FOR IDENTIFYING INTERACTION PATTERNS

The experimental setup and procedure described below was approved by the Human Research Ethics Committee at Delft University of Technology on August 17th' 2020 (reference number: 1261).

Participants

A total of 24 people participated in the experiment (17 female, seven male), recruited through personal connections on LinkedIn, within the university, from a Slack community on AI and Design and from interns at TNO. The average age among the participants was 24.8 (Std = 2.47). All of the participants had a

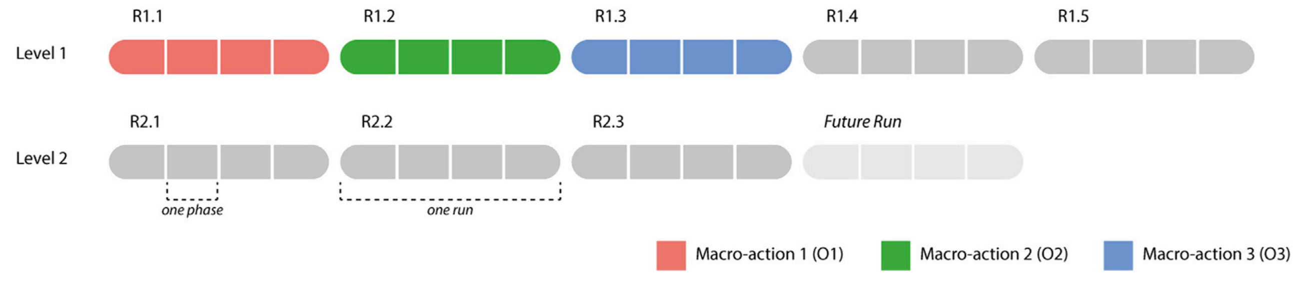


FIGURE 6 | An overview of the representation of the learning problem embedded in the experiment. It shows the different runs that a participant went through (5 runs for level 1, 3 runs for level 2), as well as how the runs were separated into 4 phases defined by the Phase Variables. The colors show how in R1.1, R1.2 and R1.3, the robot usually used *O1—picking up all*, *O2—passive large rocks* and *O3—breaking* respectively in each phase. From R1.4 onwards, the robot would choose a Macro-action based on the learned Q-values. The Future Run portrays the behavior that the robot would engage in if there were another run, based on the Q-values *after* R2.3.

university degree in a STEM field. Most of them had little to some experience with gaming (n = 7 for "little experience," n = 10 for "some experience"). Also, most of them expressed that they had no to little experience with human-robot collaboration (n = 11 for "no experience," n = 7 for "little experience") or human-robot collaboration research (n = 11 for "no experience," n = 5 for "little experience").

Due to a few problems in the data collection, some participants were excluded from all of the analyses or some of the analyses. Two participants (one female, one male) were excluded from all analyses, because there were significant connectivity issues during the execution of the experiment and/or data collection went wrong on more than one factor. One participant (female) was excluded from the questionnaire analyses, because their data was not properly saved, and one participant (male) was excluded from the robot behavioral analyses, because the log data was not properly saved.

Design and Materials

Participants were divided over two conditions: 1) a condition in which participants were instructed to think aloud and 2) a condition in which they were asked to perform the task in silence. Since we study learning processes, and since it is known that thinking aloud can have an effect on learning, the two conditions ensured that we had control over any possible effect.

We presented the task environment described in *Task Implementation* to all participants, in the form of two different levels. The first level was designed to be relatively easy, as it could be solved by simply clearing away all rocks (**Figure 2A**). A complicating factor was that breaking rocks would easily hurt the victim, which would therefore need to be avoided. The second level was designed to be more challenging: it contained a brown rock that could not be picked up at all (**Figure 2B**). This means that if the brown rock would fall on top of the victim, it would no longer be possible to save the victim and finish the task.

Participants played the first level five times, as it was estimated from pilot runs that five times would provide ample opportunity for both the participant and the robot to learn a working strategy. Participants then played the second level three times, to give the team the opportunity to adapt to the new situation. The repetition allowed for within-subject analyses, in which the behavior of participants could be compared between rounds, as well as between-subjects analyses of learning. For an overview of how the definition of the task in the Reinforcement Learning algorithm combines with this setup, **Figure 6**.

Procedure

The experiment was conducted through a video call between the experimenter and each individual participant, while both were located in their own home for the course of the experiment. Participants were given access to the experimental task using Parsec, which is a screen-sharing platform made for collaborative gaming (Parsec., n.d.). This ensured that participants had control over the task environment, while allowing the experimenter to observe their behavior.

All participants went through the following steps:

1. Participants were seated in front of their own computer at home;
2. They read the instruction, signed the consent form and provided some demographic information as well as information on their experience with video games, human-robot collaboration and human-robot collaboration research;
3. Participants had the opportunity to do a short test scenario of the task without the virtual robot, to familiarize them with the task environment and the controls;
4. Participants were presented with the first pre-specified level. After five runs, the new level was presented to the participant, which they played three times;
   a. The participants in condition A were asked to think aloud during the execution of the levels;
   b. After each level, participants completed a selection of the questionnaire on Subjective Collaboration Fluency [taken from (Hoffman 2019), see **Supplementary Appendix SA** for the questions used]. In addition, they were asked to rate how confident they were that their strategy was a good strategy for solving the task on a scale of 1–10;
   c. After the first five runs and at the end of the experiment the participants were interviewed about their experiences.

Data Collection and Analysis

Several types of data were collected in order to answer our research questions:

1. Screen captures and notes of behavior in the MATRX environment during the execution of the experiment
2. Voice recordings of the participants in condition A while they are thinking aloud during the execution of the experiment
3. Voice recordings of short interviews (see **Supplementary Appendix SA** for the questions asked)
4. Collaboration Fluency scores
5. Confidence of Strategy scores
6. Q-table as learned by the robot and log of how it changes

We will explain in more detail how this data was collected and how it relates to our research questions in the sections below.

Behavior

In order to identify what interaction patterns drive co-adaptation and co-evolution, we wanted to look at how the behavior and strategy of the team changed over time, and which interactions were used in that process. We used data types 1, 2, 3 and 6 for this. The screen captures and notes (data type 1) serve mainly as data on the human behavior, while the thinking aloud output and the interviews (data types 2 and 3) help to explain why humans behave in a certain way. The Q-tables (data type 6) serve to see what strategy the robot chose in each phase of the task. Normally, behavior of a robot driven by RL is assessed by looking at the cumulative rewards. As we are not necessarily interested in performance, but in the behavior resulting from the learning process [as prescribed in (Rahwan et al., 2019)], we chose to look at the development of the Q-tables, to understand what macro-action the robot learned to choose in each phase.

A Grounded Theory (Charmaz 2014) process was used to identify recurring adaptive behaviors from the screen captures and notes. This means that we went through a process of open coding first, while constantly writing short memos of observed patterns. After that, we collected all codes and categorized and clustered them until reaching the desired level of detail.

We will explain how the behavioral data and Q-tables were used to answer our research questions in more detail in *Results*.

Subjective Collaboration Fluency and Confidence Score

Within the task that we designed, it is quite difficult to keep track of task performance due to the possibility for large differences in strategies, as well as because the task can become unsolvable. To still keep track of how the human-robot team performed over the course of the experiment, we have chosen two measures for tracking subjective task performance: subjective collaboration fluency and confidence score (data types 4 and 5). These measures helped us to validate that our experiment setup actually allowed for learning and improvement.

For subjective collaboration fluency, we used a short version of an existing questionnaire (Hoffman 2019). To measure participants' confidence in their strategy, we asked them to rate confidence on a scale from 1 to 10 with the following question: "How confident are you that your strategy is the right strategy?"

RESULTS

Subjective Collaboration Fluency and Confidence Score

We have created a box plot of the Subjective Collaboration Fluency scores (**Figure 7**). The Confidence scores followed a very similar pattern, therefore we do not go into further detail about those. Both scores follow a pattern with scores starting off relatively high in run one, after which they drop for run two and three, move up again for four and five, drop again for six and then move up for the last two runs. To test whether thinking aloud and the number of the run affected participants' experience of collaboration with the robot, the Subjective Collaboration Fluency score was entered in a one-way repeated measures ANOVA with Thinking Aloud (yes/no) as between-subjects factor, and run number (1–8) as a within-subjects repeated measure factor. Results show that there was no significant difference between the participants who were instructed to think aloud (Mean = 40.56, Std = 22.78) and those who were not (Mean = 45.12, Std = 25.75) ($F = 0.81$, $p = 0.38$), while there was a significant effect on the run ($F = 5.97$, $p < 0.0001$). When looking at **Figure 7**, we expected this significant difference to exist between round one and round two, round three and round four, round five and round six and round six and round seven (scores went down after round one, up after round three, down again after round five and up after round six). To test whether these differences between rounds were significant, we did a post-hoc analysis using a Tukey HSD test, which mostly confirmed the differences visible from the plot: R1.1 and R1.2 are significantly different ($p = 0.006$), R1.3 and R1.4 are significantly different ($p = 0.006$), R1.5 and R2.1 are almost significantly different ($p = 0.058$) but R1.4 and R2.1 are ($p = 0.003$). R2.1 and R2.2 did not differ statistically, but R2.1 and R2.3 do, although not significantly ($p = 0.114$).

The pattern of scores on both fluency and confidence over runs are probably caused by the setup of our experiment. In the first run, the robot would use macro-action 1 for the whole task, which is the easiest to work with from a participant perspective. Therefore, participants may have been inclined to assign high scores in the beginning. In run 2 and 3, the way the RL algorithm is implemented causes the robot to use macro-action 2 and 3 respectively for the whole task, which are quite hard to understand from a participant perspective, arguably leading to a lower level of experienced fluency. In run 4 and 5, the robot would start picking its macro-action based on previous performance, while the participant would have learned to work with the robot a bit more. It is likely that this made the fluency scores go up again. Run 6, however, introduced the new scenario in which previously learned strategies often did not work anymore. In runs seven and eight the human-robot team would then learn to perform better at this second scenario, inducing participants to give higher scores on experienced collaboration fluency. Following this explanation of the scores, these results

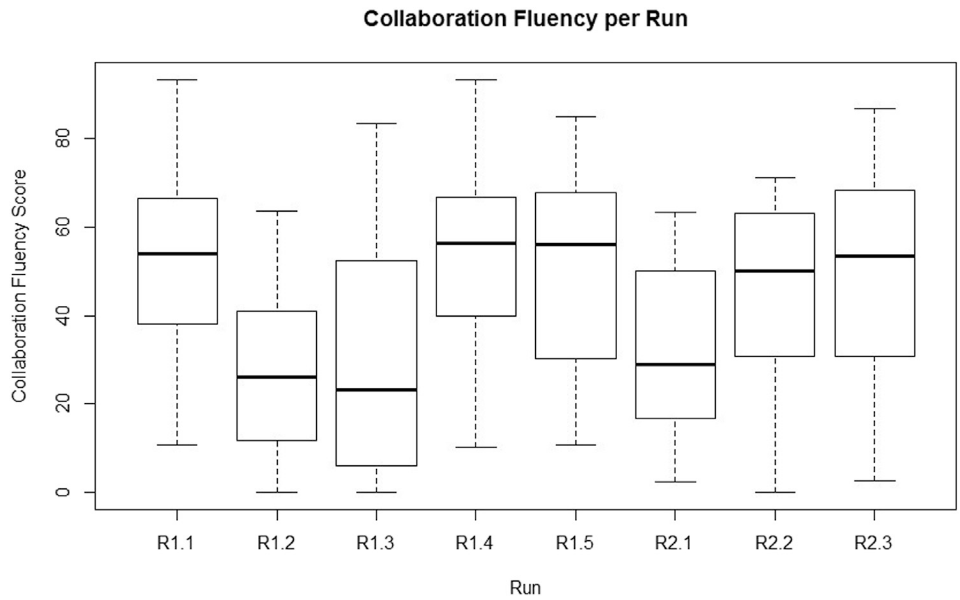


FIGURE 7 | The Collaboration Fluency scores per run in the experiment for all participants.

suggest that our experiment design indeed allowed for a learning process of the human-robot team as we anticipated.

Interaction Pattern Analysis

The open coding process of behavioral summaries, based upon the information from videos, interviews and notes, yielded a list of 52 different behaviors. These behaviors consist of *task-related actions* by the participant; *interactions* between the participant and the virtual robot; *learning* (participant learns something about the task or the collaboration); *strategies* (combinations of actions executed over longer periods of time); *team performance* and participant *emotional responses*. After excluding behaviors that did not relate to adaptation in specific (e.g. actions such as "picking up top rocks") and behaviors that were more of an assessment of the quality of a behavior rather than a description (e.g. performance factors such as "not understanding the link between waiting and robot action"), a list of 38 behaviors was left.

These 38 behaviors were categorized in the following two categories [based on the categorizations made in (van Zoelen et al., provisionally accepted)]:

- Stable situations (9 behaviors): behaviors observed *in-between* adaptations, such as the behavior of the participant alternating acting and waiting for the robot.
- Sudden adaptations (29 behaviors): behaviors in which the human and/or robot adapted their actions, thus starting a transition from one stable situation to another. The adaptation happens in a single moment or over a short period of time, often in response to a newly hypothesized or discovered property of the partner's behavior.

The full list and categorization can be found in **Supplementary Appendix SB**. The behaviors listed in the Appendix are closely tied to the experimental task. The descriptions of the behaviors were processed to fit co-adaptation in general (such that we can call them interaction patterns). For this purpose, some of the behaviors were combined into one descriptive interaction pattern. The resulting list, consisting of 23 interaction patterns (five stable situations, 18 sudden adaptations), is presented in **Table 3**.

The biggest group is that of "sudden adaptations," the patterns that often arise in response to a discovery, an expectation, or a surprise of one of the partners. In order to better understand this important group of adaptive interaction patterns, we explored in more detail the nature of the triggers that initiate them, what characterizes the execution of these patterns, and what they bring about in the human-robot collaboration. Again following the approach taken in [van Zoelen et al. (provisionally accepted)], we used the following terms to describe the sudden adaptations:

- External trigger: an event outside of the partner (e.g. in the task, environment or other partner) triggers an adaptation to a new stable situation;
- Internal trigger: an event inside of the partner (e.g. a specific expectation or change of mind) triggers an adaptation to a new stable situation;
- Outcome: a specific action that is preceded by an internal or external trigger, that will gradually develop into a new stable situation afterward;
- In-between-situation: a specific action that is preceded by an internal or external trigger, that serves as a new trigger for adapting to a new stable situation afterward.

The results of this can be found in **Supplementary Appendix SB**.

TABLE 3 | The interaction patterns identified from the behavioral data, including a description of what they entail.

Category	Concept	Description
Stable situation	Actively synchronizing actions with a team member	Human understands the capabilities of another team member and actively uses their own actions to make optimal use of the combined capabilities
	Alternating actively working on the task and waiting for a team member	Human switches between performing their own task for a while, then waiting for a team member to perform their task, and so on
	Being generally passive and letting a team member do most of the work	Human is overall passive and lets the other team member do the work
	Damage control: Prevent damage caused by a team member	Human performs actions that prevent their team member from causing intentional or unintentional harm or damage
	Focusing on own task	Human performs their own task without paying much attention to their team member
Sudden adaptation	Avoiding communication with a team member	One of the team member actively avoids the other team member to avoid unwanted communication interpretations
	Being confused by non-human-like behavior	A human team member is confused by non-human-like behavior performed by a team member
	Being confused by unexpected behavior (negative)	One of the team members is confused or frustrated by behavior performed by their team member that they did not expect
	Being happy that a team member does as expected	One of the team members is happy that their team member performs the kind of behavior that they expect and hoped for
	Being surprised by unexpected behavior (positive)	One of the team members is positively surprised by behavior performed by their team member that they did not expect
	Coming into action when a team member comes into action	A team member starts to actively perform their task after a period of inaction, when their team member also starts to actively perform their task after a period of inaction
	Doing useless or harmful actions because there is nothing else to do	A team member is unable to perform useful actions, therefore starts performing useless or harmful actions
	Feeling alone, as if team member does not help	A human team member feels left alone
	Following a team member's action	A team member follows or copies the action performed by another team member
	Learning about behavioral cues	A team member gains insight into specific behavior performed by another team member
	Learning about own capabilities	A team member gains insight into their own capabilities
	Learning about team member's capabilities or strategy	A team member gains insight into the capabilities or strategy of another team member
	Moving around different task components	A team member moves around different task components without actually performing any task
	Team member changes strategy, which is visible by a behavioral cue	A team member observes that another team member changes strategy by a behavioral cue
	Team member performs an action that makes no sense	A team member performs a useless action
	Trying to communicate by interacting with a team partner	A team member attempts to communicate with another team member by directly interacting with them, for example by coming close to them
	Trying to communicate by signaling task actions	A team member attempts to communicate with another team member by trying out different actions that they want their team member to perform
	Waiting for a team member to start acting	A team member waits for another team member to start performing their task

Collaborative Learning

In addition to developing a comprehensive description of adaptive interaction patterns, we further explored how human behavior, and specifically human behavior adaptation, influenced learning by the virtual robot. We analyzed and coded human adaptive behavior at a detailed level by identifying the interaction patterns as described above, but in order to analyze how the development of robot behavior and human learning depend on each other, a different level of detail was necessary. We looked at three aspects of the data:

- For each participant, we looked at the Q-tables of the virtual robot at the end of each run in the experiment, to see which of the three macro-actions received the highest expected reward in the different phases and runs of the task;
- For each participant, we identified the main behavioral strategy used by the human per run, as well as during the whole experiment;
- We analyzed how the chosen macro-actions of the virtual robot can be associated with specific behavioral strategies of the participants.

We will describe our process for all three of these aspects in more detail below.

Virtual Robot Q-Tables

The Reinforcement Learning algorithm addressed learning when to apply which of three macro-actions or options (as described in **Figure 3**, **Figure 4**, and **Figure 5**; we will call them *O1—picking up all*, *O2—passive large rocks* and *O3—breaking* from here onwards) and used four phase variables to identify states. **Figure 8** shows an overview of how often the robot learned to pick specific macro-actions in each phase (**Figure 8A**) and each run (**Figure 8B**), based on the macro-action with the highest expected reward.

As the robot's choice for macro-options is not clearly related to the phases in the task (especially phase 2, 3 and 4 are very similar, as can be seen in **Figure 8A**), we looked mostly at **Figure 8B** to understand how the robot's behavior developed. The figure shows that in the first three runs the robot mostly tried out all macro-strategies one by one, as determined by how the algorithm was programmed. Small deviations from this are likely caused by some participants going back and forth between phases in the

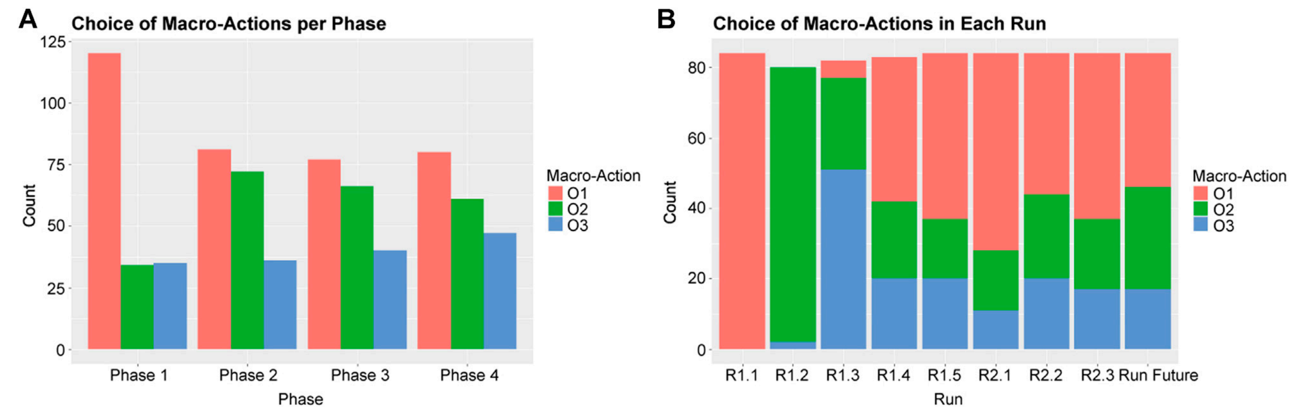


FIGURE 8 | An overview of how often certain Macro-actions were chosen by the robot across all participants per phase **(A)** and per run **(B)**.

task, rather than moving through them linearly as we initially expected. The robot generally learned to select *O1—picking up all* most of the time for most participants over the course of the next few runs, which fits with how level 1 of the experiment was designed. From run 6 onwards, when the second level was introduced, the robot learned to choose *O2—passive large rocks* and *O3—breaking* more often. This shows that the robot is able to generally learn what works best for the task.

Participant Behavioral Clustering

To better understand how the behavior of the participants developed over time, we performed a manual clustering of participant behavior per run. Based on the behavior observations as described in *Interaction Pattern Analysis*, we defined the following behavioral clusters:

- Just focus on own behavior efficiently
- Balancing acting and waiting
- Exploring how the robot works by observing and trying to communicate
- Actively using *O3—breaking*
- Actively using *O2—passive large rocks*

The result of this clustering can be seen in **Figure 9**. It is difficult to find detailed insights from this figure, apart from the fact that more participants showed more adaptive behavior in the later runs, as indicated by the red and olive green bars in the figure. Participants' strategies did not develop linearly, and it also did not converge to one specific type of behavior consistently within our experiment. To be able to see whether human learning had an influence on robot learning, we chose to remove the dimension of time (runs) from our participant data, and focus on whether and how a participant adapted over the whole experiment. We created the following clusters based on participant adaptation over the whole experiment:

- Does not adapt: participant shows no signs of adapting to strategies employed by the robot; participant either focuses on their own task, or constantly switches between behavior strategies as they focus too much on the robot.
- Adapts by balancing waiting and acting: participant shows signs that they adapt by waiting for the robot to act, and to use that robot behavior to determine their own response. It suggests that the participant understands that being passive for a while may cause the robot to act.
- Adapts by actively using *O2—passive large rocks* or *O3—breaking*: participant visibly adapts as they actively guide the robot to pick up or break rocks by waiting on top of those rocks.

This clustering of participants according to their dominant strategy resulted in three clusters with a similar number of participants per cluster, as shown in **Table 4**.

Combining Participant Adaptation and Robot Learning

In order to explore whether these different types of adaptation employed by participants affected robot learning, and whether differences occur between clusters, we plotted the robot strategies per human adaptation cluster, as shown in **Figure 10**. These figures present the bar graphs of how often the different macro-actions were chosen across the group, in an attempt to more closely evaluate any possible differences between the three clusters. The figures suggest a response in the robots' behavior to the participants' actions, especially later on, in the final runs. When we compare the figures for "adapts by balancing passively waiting and acting" (**Figure 10B**) and "does not adapt" (**Figure 10C**), the former shows the trend of using *O1—picking up all* in the first level (first five runs) and using other strategies in the second level (run 6–8) more strongly. The figures for "adapts by actively using O2 or O3" (**Figure 10A**), however, shows that the robot already learned to use *O2—passive large rocks* and *O3—breaking* before being introduced with the second level, and actually moving back to *O1—picking up all* a little more toward the final runs.

This suggests that if participants learned more about the robot behavior and adapted their own behavior more strongly (by actively guiding the robot with *O2—passive large rocks* and *O3—breaking*), the robot was also able to learn to use those strategies more often. This combined learning effect therefore can be seen as learning at the team level.

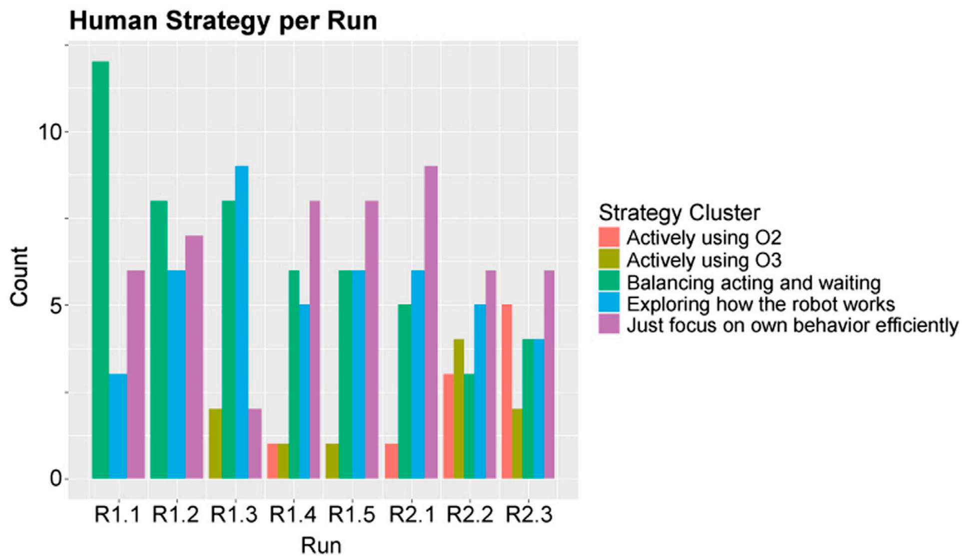


FIGURE 9 | An overview of how many participants used specific behavioral strategies per run.

TABLE 4 | The clusters resulting from manually clustering participants based on whether they adapted to the robot across the whole experiment.

Cluster	Participants
Does not adapt	2, 6, 9, 15, 21, 22, 23, 24 (n = 8)
Adapts by balancing passively waiting and acting	12, 13, 14, 16, 27, 28 (n = 6)
Adapts by actively using O2 or O3	3, 8, 10, 17, 19, 20, 26 (n = 7)

DISCUSSION

Interaction Patterns That Drive Co-Learning

We set out to investigate what interaction patterns between humans and robots drive co-adaptation as a precursor for co-learning. From the behaviors observed in our experiment, we identified a list of interactions and a set of interaction patterns. It should be noted, however, that (interactive) behavior is very much determined by the specific context. This means that our list should not be considered as a complete list of all possible co-adaptive interactions. It should rather be seen as a collection of interaction patterns that are likely to appear in contexts similar to ours, where co-adaptation is centered around mutual observation and harmonizing actions when collaborating.

The collection of interaction patterns can be used as a language for recording, analyzing and coding co-adaptive behavior. By describing observed behavior with such interaction patterns, complex behavioral observations of human-robot co-adaptive strategies can more easily be compared. The interaction patterns can also be useful as a vocabulary for a human-robot team itself to discuss the adaptations that they are engaged in, to help them elaborate and sustain successful collaborations over time.

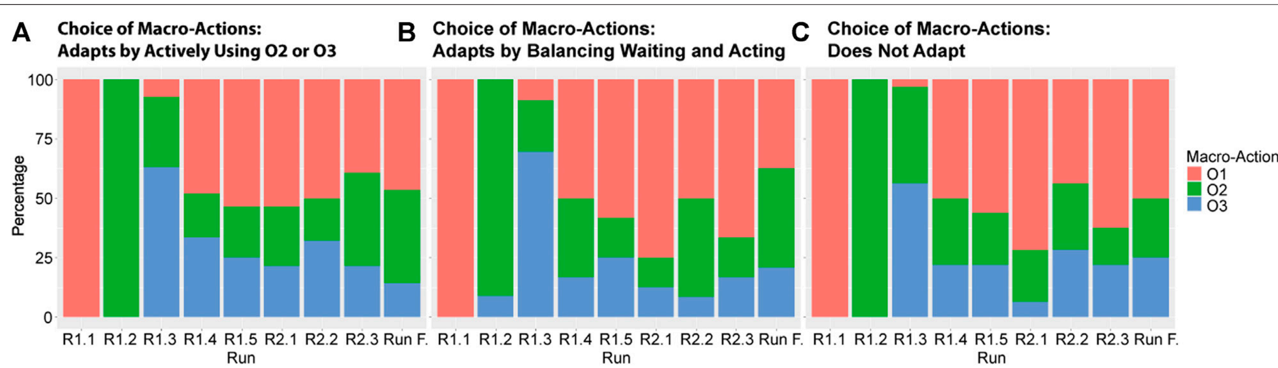


FIGURE 10 | An overview of how often certain Macro-actions were chosen by the robot across all participants per run, split up by the level adaptation the participant showed: **(A)** shows participants who adapted by actively using *O2—passive large rocks* and/or *O3—breaking*, **(B)** shows participants who adapted by balancing waiting and acting, and **(C)** shows participants who did not adapt.

TABLE 5 | The claims as presented in *Claims: Expected Observations* that need to be justified in a co-adaptation experiment, including whether they were validated and an explanation of that conclusion.

Claim	Justified	Explanation
Different participants develop different ways of performing the task	Yes	When looking at the different interaction patterns that people engage in, and categorizations of their adaptive behavior, we can see that different people indeed performed the task in a variety of ways
The agent learns different sequences of strategy options for different participants	Partly	The results showed that not all agents learned the same model on an individual level. However, the models had much in common, suggesting that all agents learned similar behavior. When splitting this up in groups based on human adaptive behavior, there seems to be a difference in learned agent behavior between the different groups. Currently, however, we did not do any statistical analysis to test whether this is a significant result
Different teams converge to different ways of performing the task	Partly	When looking at the different interaction patterns that participants engaged in with their robot team partner, different teams solved the task in a variety of ways (see H1). However, it is unclear to what extent the robot contributed to this. Moreover, while participants generally gained more confidence in their strategy and expressed to experience a greater subjective collaboration fluency toward the end of the experiment, it is unclear to what extent the strategy of the team really converged to a stable one
The agent converges to a specific sequence of strategy options for most participants	No	While we did observe a logical development of the Q-values on a population level, this does not count for all of the individual agents. Moreover, it is not clear to what extent the agents really converged to a stable set of actions
The human converges to a specific strategy within the experiment	Partly	The categorizations of participant behavior show that participants settle on a stable strategy more and more over the course of the experiment. This is also shown by the development of the confidence scores and subjective collaboration fluency. True convergence to a stable strategy, however, is not clearly visible within the 8 runs of the experiment

Validating the Research Environment

An important objective of this study is to improve our understanding of how human-robot co-learning develops, as well as how the adaptative processes of both partners interact. We defined claims for the experimental environment we designed; if these claims are justified, it means that it enabled us to study co-adaptation, the process that we consider to be a precursor for co-learning. **Table 5** shows the claims described in *Claims: Expected Observations* and annotated conclusions as to whether we were able to justify the claims in the present study.

As can be seen in the table, only one of the claims was justified completely. Fortunately, many of the other claims were partly justified. For the claims that we did not realize, the results provide cues for how to design a research environment that better fits the claims. We regard these findings as an important step toward studying and revealing the processes involved in human-robot co-learning. Aspects that should be improved upon or need further work center around a few problems that we will elaborate on below:

1. Behavior strategy: Convergence vs. flexible adaptations
2. Statistical analysis of complex behavioral data
3. Behavior of the individual team member vs. behavior of the team
4. Task effects vs. participant effects

In our claims in Table 7, we mentioned convergence several times. This stems from the principle that a Reinforcement Learning algorithm should aim for convergence toward an optimal solution. However, when studying co-learning, we specifically use dynamic task environments that have no fixed optimal solution, and in which unpredicted events can require strategy changes. In such environments, convergence is not a good criterion for performance, as agents (human as well as robot) are required to continuously learn and adapt. For human-robot co-learning it can be argued that it is better to make the algorithm learn certain repeated subsequences of interactions (or *interaction patterns*), and to store those in a rule-based manner. Once a pattern of interaction has proven to be successful in multiple instances of task situations, it can be applied, combined and if necessary revised in similar but other task situations. We therefore believe that future research into co-learning should not take convergence as a criterion for the robot's behavior, but to focus on the emergence and sustainability of successful interaction patterns (aspect 1).

The results show that the robot had a similar learning process across all participants despite the high variety between individual participants. However, the behavioral data of both the robot and the participant is quite complex. Sometimes there are radical changes in behavior between one run and the next, and even within one run participants sometimes quite radically changed their behavior. It is a challenge to analyze such data as it is often difficult to clarify the origin of the behavior from the data. Our qualitative analysis and clustering is able to deal with this complexity and provides many useful insights, therefore we would advise future research into co-learning to include similar qualitative analyses. When further investigating co-learning, it will, however, also be relevant and interesting to verify insights statistically. This will require different design considerations. The current complexity in behavioral data is partly due to the interaction between two adaptive systems, and probably an inherent property of co-learning. Moreover, the human and robot can approach the task in many different ways by design. This property is a strength of our experiment, as it allows participants to behave relatively freely and naturally, but it also contributes to the complexity of the resulting co-adaptation. For future research, it will be important to explicitly take these properties into account when designing an environment for experimentation (aspect 2), in such a way that insights can be verified statistically. For example, in our design, the strategies of the robot were separate, nominal actions, but if we can design a learning agent such that their learned behavior is ordinal (e.g. by using more or less of a certain behavior), it might be easier to apply statistical methods. Moreover, we can look into data analysis methods used in complexity science, to see if they can be applicable to co-learning scenarios.

Lastly, it is currently a challenge to determine which aspects of the final team behavior are caused by adaptations of individual team members, and which by interactions between them. Similarly, we cannot yet conclusively determine which aspects of the learned strategy are caused by the task, and which by the individuality of a participant (e.g. what does the robot learn just because of the task, and what does the robot learn because a certain participant behaved a certain way). To solve this problem, we need to find ways to separate the different effects, for example by creating relevant baseline results. Letting the robot perform a task and learn by itself is not an option in the context of team tasks, as the nature of such task dictates dependencies between the team members. A possibility might be to create a simulated, possibly rule-based human agent for the robot agent to collaborate with.

Future Steps for Studying Human-Robot Co-Learning

The discussion regarding the interactions that underlie co-learning provide several pointers and suggestions for improving the design of our research methods. Besides the suggestions above, however, there are several other directions in which we believe that co-learning research should develop. In this paper we developed an approach for studying co-learning. Since we were still defining what it means to study co-learning, the scope of the task, learning algorithm and opportunities for interaction had to be limited. Eventually, if we want to enable and study co-learning in full-fledged teams, it will be necessary to use more complex task environments, more intelligent agents or robots and more elaborate interaction and communication between the human and the robot. In the following section we will therefore further outline the two research directions mentioned in *Research Challenges* that we did not further address in this paper (research direction two and three).

The first direction is aimed at enabling the communication between partners within the team, especially communication about adaptations. We believe that in order for team members to produce successful sequences of interactive behavior, that can be used strategically across contexts, it is necessary that the team members can communicate with one another. The interaction patterns that we have identified might be used as a start for a vocabulary for such communication interactions, but the specific timing, modality and details of the interactions will have to be designed and studied.

The second direction is aimed at making the agent or robot more intelligent in terms of its abilities for co-learning. In the experiment we presented, our agent only learned based on task-related rewards. It makes sense to also explicitly reason about or take into account the human's behavior and preferences in learning. There is a large body of research on personalizing robot behavior, e.g. by making the robot develop a user model of its partner, but these models often do not explicitly take into account that the human continuously learns and adapts. We therefore believe that there is a need for user models and team models that specifically accommodate the adaptive interactions as described in this paper. A team mental model that is able to represent the interactions within a team will support the partners in developing and sustaining successful adaptations and to synchronize and align their actions and learning processes.

CONCLUSION

Co-learning is an important mechanism for building successful human-robot teams. However, there is no general understanding of what co-learning means in the context of human-robot teams. In this paper, we defined the concept of co-learning based on related literature, and positioned it in relation to co-adaptation and co-evolution. From this definition, it is clear that adaptive interactions between humans and robots play a central role in co-learning. We defined requirements for studying how bi-lateral adaptation emerges from the interactions between humans and robots.

From these requirements, we developed an experimental task environment based on a real-life Urban Search and Rescue task. The task was designed such that it allowed human participants to behave relatively naturally and freely, enabling us to record and analyze emerging adaptive interactions between a human and a robot. A bottom-up coding process based on Grounded Theory allowed us to identify recurring interaction patterns. The resulting list of interaction patterns describe stable situations (repeating subsequences of stable behavior) as well as sudden adaptations (changes in behavior happening over short periods of time). These patterns can emerge in similar task contexts, thereby forming a description and analysis method of co-adaptive behavior in human-robot teams.

Over the course of the experiment, the robot learned similar strategies for most participants. However, the results show that the learned strategies were slightly different depending on whether a human participant adapted their behavior to the robot. This suggests that human learning affected the robot's Reinforcement Learning. More specifically, the human learning about and adapting to a specific strategy of the robot enabled the robot to learn to stick to that strategy. This shows how learning on the individual level lead to team level learning in our experiment.

This paper presents a theoretical framework and a methodological approach for studying the processes that underlie co-learning in human-robot teams. The strengths as well as the shortcomings of our approach provide ample directions for future research into this important process that ultimately defines the quality of human-robot teaming.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the following online repository: van Zoelen, Emma (2021), "Mutual Adaptation for Human-Robot Co-Learning - USAR task", Mendeley Data, V1, doi:10.17632/r2y8z6bzg8.1.

ETHICS STATEMENT

The study involving human participants was reviewed and approved by the Human Research Ethics Committee at Delft University of Technology (reference number: 1261). The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The main research design, experimentation and analysis as well as most of the writing was done by EV. All of this was done under supervision of KV and MN. Both KV and MN also reviewed and edited the paper at different stages of the process.

FUNDING

This study was a collaboration between the Netherlands Organisation of Applied Scientific Research (TNO) and Delft University of Technology, funded by TNO under contract "4099 DO AIO Fonds," and partially carried out within the APPL.AI FATE project and the AIMS projects (V1801).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.692811/full#supplementary-material

REFERENCES

Ansari, F., Erol, S., and Sihn, W. (2018). Rethinking Human-Machine Learning in Industry 4.0: How Does the Paradigm Shift Treat the Role of Human Learning? *Proced. Manufacturing* 23, 117–122. doi:10.1016/j.promfg.2018.04.003

Baxter, P., Ashurst, E., Read, R., Kennedy, J., and Belpaeme, T. (2017). Robot Education Peers in a Situated Primary School Study: Personalisation Promotes Child Learning. *PLoS ONE* 12 (5), e0178126. doi:10.1371/journal.pone.0178126

Belpaeme, T., Vogt, P., van den Berghe, R., Bergmann, K., Göksun, T., de Haas, M., et al. (2018). Guidelines for Designing Social Robots as Second Language Tutors. *Int. J. Soc. Robotics* 10 (3), 325–341. doi:10.1007/s12369-018-0467-6

Bosch, K. v. d., Schoonderwoerd, T., Blankendaal, R., and Neerincx, M. (2019). Six Challenges for Human-AI Co-learning. *Lecture Notes Comput. Sci.* 11597, 572–589. doi:10.1007/978-3-030-22341-0_45

Burke, C. S., Stagl, K. C., Salas, E., Pierce, L., and Kendall, D. (2006). Understanding Team Adaptation: A Conceptual Analysis and Model. *J. Appl. Psychol.* 91 (December), 1189–1207. doi:10.1037/0021-9010.91.6.1189

Buschmeier, H., and Kopp., S. (2013). Co-constructing Grounded Symbols-Feedback and Incremental Adaptation in Human-Agent Dialogue. *Künstl Intell.* 27, 137–143. doi:10.1007/s13218-013-0241-8

Charmaz, K. (2014). *Constructing Grounded Theory*. London, United Kingdom: SAGE.

Chauncey, K., Harriott, C., Prasov, Z., and Cunha, M. (2016). "A Framework for Co-adaptive Human-Robot Interaction Metrics," in Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, Daejeon, Korea, October 9–14, 2016 (New York, NY, USA: ACM), 123–124.

Dillenbourg, P., Baker, M., Blaye, A., and O'Malley, C. (1996). The Evolution of Research on Collaborative Learning. *Learning in Humans and Machine: Towards an Interdisciplinary Learning Science*. Oxford: Elsevier, 189–211.

Döppner, D. A., Derckx, P., and Schoder, D. (2019). "Symbiotic Co-evolution in Collaborative Human-Machine Decision Making: Exploration of a Multi-Year Design Science Research Project in the Air Cargo Industry," in Proceedings of the 52nd Hawaii International Conference on Systems Science, Grand Wailea, Hawaii, January 8–11, 2019, 264–273. doi:10.24251/HICSS.2019.033

Dulac-Arnold, G., Mankowitz, D., and Todd, H. (2019). Challenges of Real-World Reinforcement Learning. ArXiv:1904.12901 [Cs, Stat], April. http://arxiv.org/abs/1904.12901.

Ehrlich, S. K., and Cheng, G. (2018). Human-Agent Co-adaptation Using Error-Related Potentials. *J. Neural Eng.* 15 (6), 066014. doi:10.1088/1741-2552/aae069

Foerster, J., Assael, I. A., de Freitas, N., and Whiteson, S. (2016). "Learning to Communicate with Deep Multi-Agent Reinforcement Learning," in *Advances in Neural Information Processing Systems 29*. Editors D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Barcelona, Spain: Curran Associates, Inc), 2137–2145. http://papers.nips.cc/paper/6042-learning-to-communicate-with-deep-multi-agent-reinforcement-learning.pdf.

Gao, Y., Barendregt, W., and Castellano, G. (2017). Personalised Human-Robot Co-adaptation in Instructional Settings Using Reinforcement Learning. Proceedings of the Persuasive Embodied Agents for Behavior Change (PEACH2017) Workshop, Uppsala, Sweden, August 27–30, 2017. Stockholm, Sweden.

Hoffman, G. (2019). Evaluating Fluency in Human-Robot Collaboration. *IEEE Trans. Human-mach. Syst.* 49 (3), 209–218. doi:10.1109/THMS.2019.2904558

Holstein, K., Aleven, V., and Rummel, N. (2020). A Conceptual Framework for Human-AI Hybrid Adaptivity in Education. *Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán, 240–54. Lecture Notes in Computer Science*. Cham: Springer International Publishing, 240–254. doi:10.1007/978-3-030-52237-7_20

Illanes, L., Yan, Xi., Toro Icarte, R., and Sheila, A. M. I. (2019). Symbolic Planning and Model-free Reinforcement Learning - Training Taskable Agents. https://drive.google.com/file/d/15b4w56ODJa4PZpSNbhho5q1s3pl_DlyP/view?usp=embed_facebook.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, M. B., and Sierhuis, M. (2014). Coactive Design: Designing Support for Interdependence in Joint Activity. *J. Human-Robot Interaction* 3 (1), 43. doi:10.5898/JHRI.3.110.5898/jhri.3.1.johnson

Kapetanakis, S., and Kudenko, D. (2002). "Reinforcement Learning of Coordination in Cooperative Multi-Agent Systems," in Proceedings of the Nineteenth National Conference on Artificial Intelligence, Edmonton, AB, July 28–August 1, 2002 (Edmonton), 326–331.

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., and Feltovich, P. J. (2004). Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intell. Syst.* 19 (6), 91–95. doi:10.1109/MIS.2004.74

Lee, C.-S., Wang, M.-H., Huang, T.-X., Chen, L.-C., Huang, Y.-C., Yang, S.-C., Tseng, C.-H., Hung, P.-H., and Kubota, N. (2018). "Ontology-Based Fuzzy Markup Language Agent for Student and Robot Co-learning," in 2018 IEEE

International Conference on Fuzzy Systems, Rio de Janeiro, Brazil, July 8–13, 2018 (FUZZ-IEEE), 8–13. doi:10.1109/FUZZ-IEEE.2018.8491610

Lee, C.-S., Wang, M.-H., Tsai, Y.-L., Ko, L.-W., Tsai, B.-Y., Hung, P.-H., et al. (2020). Intelligent Agent for Real-World Applications on Robotic Edutainment and Humanized Co-learning. *J. Ambient Intell. Hum. Comput* 11 (8), 3121–3139. doi:10.1007/s12652-019-01454-4

Lematta, G. J., Coleman, P. B., Bhatti, S. A., Chiou, E. K., McNeese, N. J., Demir, M., and Cooke, N. J. (2019). "Developing Human-Robot Team Interdependence in a Synthetic Task Environment," in Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Seattle, WA, October 28–November 1, 1503–1507. doi:10.1177/1071181319631433631

MATRX Software (2020). Python. MATRX Software. Available at: https://github.com/matrx-software/matrx.2021

Mohammad, Y., and Nishida, Toyoaki. (2008). Human Adaptation to a Miniature Robot: Precursors of Mutual Adaptation. RO-*MAN 2008 - the 17th IEEE International Symposium on Robot and Human Interactive Communication* (Munich, Germany: IEEE), 124–129. doi:10.1109/ROMAN.2008.4600654

Nikolaidis, S., Hsu, D., and Srinivasa, S. (2017a). Human-Robot Mutual Adaptation in Collaborative Tasks: Models and Experiments. *Int. J. Robotics Res.* 36, 618–634. doi:10.1177/0278364917690593

Nikolaidis, S., Zhu, Y. X., Hsu, D., and Srinivasa, S. (2017b). "Human-Robot Mutual Adaptation in Shared Autonomy," in Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17, Vienna, Austria, March 6–9 (Vienna, Austria: ACM Press), 294–302. doi:10.1145/2909824.3020252

Parsec. (n.d.). Parsec Gaming. Avilabel at: https://parsec.app (Accessed February 12, 2021).

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., et al. (2019). Machine Behaviour. *Nature* 568, 477–486. doi:10.1038/s41586-019-1138-y

Ramakrishnan, R., Zhang, C., and Shah, J. (2017). Perturbation Training for Human-Robot Teams. *jair* 59 (July), 495–541. doi:10.1613/jair.5390

Shafti, Ali., Jonas, T., Dudley, W., and Aldo Faisal, A. (2020). Real-World Human-Robot Collaborative Reinforcement Learning. ArXiv:2003.01156 [Cs], July. http://arxiv.org/abs/2003.01156. doi:10.1109/iros45743.2020.9341473

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., et al. (2015). A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. ArXiv:1506.06714 [Cs], June. Available at: http://arxiv.org/abs/1506.06714. doi:10.3115/v1/n15-1020

Stolle, M., and Precup, D. (2002). "Learning Options in Reinforcement Learning," in *Lecture Notes in Computer Science*. Editors S Koenig and R C. Holte (Berlin, Heidelberg: Springer), 212–223. doi:10.1007/3-540-45622-8_16

Sutton, R. S., Precup, D., and Singh., S. (1999). Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artif. Intelligence* 112 (1–2), 181–211. doi:10.1016/S0004-3702(99)00052-1

van Otterlo, M., and Wiering, M. (2012). "Reinforcement Learning and Markov Decision Processes," in Reinforcement Learning *Marco Wiering and Martijn Van Otterlo, 12:3–42. Adaptation, Learning, and Optimization* (Berlin, Heidelberg: Springer Berlin Heidelberg). doi:10.1007/978-3-642-27645-3_1

Van Zoelen, E. M., van den Bosch, K., Rauterberg, M., Barakova, E. I., and Neerincx, M. (provisionally accepted). Identifying Interaction Patterns of Tangible Co-adaptations in Human-Robot Team Behaviors. *Fron. Psychol.* doi:10.3389/fpsyg.2021.645545

Vignolo, A., Powell, H., McEllin, L., Rea, F., Sciutti, A., and Michael, J. (2019). "An Adaptive Robot Teacher Boosts a Human Partner's Learning Performance in Joint Action," in 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). doi:10.1109/RO-MAN46459.2019.8956455

Weber, K., Ritschel, H., Lingenfelser, F., and André, E. (2018). "Real-Time Adaptation of a Robotic Joke Teller Based on Human Social Signals," in Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018 (IFAAMAS), 3.

Wenskovitch, J., and North, C. (2020). Interactive AI: Designing for the 'Two Black Boxes' Problem. *Hybrid Human-Artificial Intelligence Special Issue* (Washington, United States: IEEE Computer Society), 1–10.

Xu, Y., Ohmoto, Y., Okada, S., Ueda, K., Komatsu, T., Okadome, T., et al. (2012). Formation Conditions of Mutual Adaptation in Human-Agent Collaborative Interaction. *Appl. Intell.* 36 (1), 208–228. doi:10.1007/s10489-010-0255-y

Yamada, S., and Yamaguchi, T. (2002). Mutual Adaptation to Mind Mapping in Human-Agent Interaction. Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication (Berlin, Germany: IEEE), 105–110. doi:10.1109/ROMAN.2002.1045606

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 in Robotics and AI

ORIGINAL RESEARCH
published: 24 August 2021
doi: 10.3389/frobt.2021.676248

Individual Differences in Children's (Language) Learning Skills Moderate Effects of Robot-Assisted Second Language Learning

Rianne van den Berghe[1,2]*, *Ora Oudgenoeg-Paz*[1], *Josje Verhagen*[3], *Susanne Brouwer*[4], *Mirjam de Haas*[5], *Jan de Wit*[6], *Bram Willemsen*[7], *Paul Vogt*[5,8], *Emiel Krahmer*[6] and *Paul Leseman*[1]

[1]Department of Development of Youth and Education in Diverse Societies, Utrecht University, Utrecht, Netherlands, [2]Section Leadership in Education and Development, Windesheim University of Applied Sciences, Almere, Netherlands, [3]Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, Netherlands, [4]Department of Modern Languages and Cultures, Radboud University, Nijmegen, Netherlands, [5]Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, Netherlands, [6]Department of Communication and Cognition, Tilburg University, Tilburg, Netherlands, [7]Department of Intelligent Systems, KTH Royal Institute of Technology, Stockholm, Sweden, [8]School of Communication, Media and IT, Hanze University of Applied Sciences, Groningen, Netherlands

OPEN ACCESS

Edited by:
Wafa Johal,
University of New South Wales,
Australia

Reviewed by:
Anara Sandygulova,
Nazarbayev University, Kazakhstan
Samuel Spaulding,
Massachusetts Institute of Technology, United States
Sooyeon Jeong,
Massachusetts Institute of Technology, United States

***Correspondence:**
Rianne van den Berghe
maj.vanden.berghe@windesheim.nl

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: 04 March 2021
Accepted: 29 July 2021
Published: 24 August 2021

Citation:
van den Berghe R, Oudgenoeg-Paz O, Verhagen J, Brouwer S, de Haas M, de Wit J, Willemsen B, Vogt P, Krahmer E and Leseman P (2021) Individual Differences in Children's (Language) Learning Skills Moderate Effects of Robot-Assisted Second Language Learning. Front. Robot. AI 8:676248. doi: 10.3389/frobt.2021.676248

The current study investigated how individual differences among children affect the added value of social robots for teaching second language (L2) vocabulary to young children. Specifically, we investigated the moderating role of three individual child characteristics deemed relevant for language learning: first language (L1) vocabulary knowledge, phonological memory, and selective attention. We expected children low in these abilities to particularly benefit from being assisted by a robot in a vocabulary training. An L2 English vocabulary training intervention consisting of seven sessions was administered to 193 monolingual Dutch five-year-old children over a three- to four-week period. Children were randomly assigned to one of three experimental conditions: 1) a tablet only, 2) a tablet and a robot that used deictic (pointing) gestures (the no-iconic-gestures condition), or 3) a tablet and a robot that used both deictic and iconic gestures (i.e., gestures depicting the target word; the iconic-gestures condition). There also was a control condition in which children did not receive a vocabulary training, but played dancing games with the robot. L2 word knowledge was measured directly after the training and two to four weeks later. In these post-tests, children in the experimental conditions outperformed children in the control condition on word knowledge, but there were no differences between the three experimental conditions. Several moderation effects were found. The robot's presence particularly benefited children with larger L1 vocabularies or poorer phonological memory, while children with smaller L1 vocabularies or better phonological memory performed better in the tablet-only condition. Children with larger L1 vocabularies and better phonological memory performed better in the no-iconic-gestures condition than in the iconic-gestures condition, while children with better selective attention performed better in the iconic-gestures condition than the no-iconic-gestures condition. Together, the results showed that the effects of the robot and its gestures differ across children, which should be taken into account when designing and evaluating robot-assisted L2 teaching interventions.

Keywords: social robots, second language learning, child-robot interaction, individual differences, (language) learning skills

INTRODUCTION

The current study addresses the use of social robots in language education. Specifically, we investigated how individual differences among children affect the added value of social robots for teaching second language (L2) vocabulary to young children. While studying the effects of robots is in itself important in view of applications in education, it is crucial to compare the effectiveness of robots to that of cheaper and more accessible technological aids such as tablets. Several potential advantages of robots relative to other technologies such as tablets have been identified in extant research. For example, social robots allow for interactions that make use of the physical environment (e.g., acting upon objects, enacting particular movements or operations, using various types of gestures) and they can stimulate more natural, human-like interactions because of their humanoid appearance (Belpaeme et al., 2018; van den Berghe et al., 2019). The use of iconic gestures is known to support L2 vocabulary learning (Tellier, 2008; Macedonia et al., 2011; Rowe et al., 2013), and a robot's iconic gestures and other non-verbal cues have been found to benefit learners as well (Kory Westlund et al., 2017; de Wit et al., 2018).

Current evidence on the effectiveness of robot-assisted language learning (RALL), however, is mixed (see for reviews Kanero et al., 2018; van den Berghe et al., 2019), and there is inconclusive evidence on the possible benefits of robots over other forms of technology (Han et al., 2008; Hyun et al., 2008; Leyzberg et al., 2012; Gordon et al., 2015; Kory Westlund et al., 2015; Zhexenova et al., 2020). Specific for word learning, positive effects of robots on learning were found in several single-session studies (Tanaka and Matsuzoe, 2012; de Wit et al., 2018), while only moderate learning gains were found in multiple-session studies (Kanda et al., 2004; Gordon et al., 2016). This effect is in contrast with evidence regarding effective and impactful vocabulary training programs involving human tutors where multiple sessions with a large number of repeatedly presented words are usually more effective than single sessions (Marulis and Neuman, 2010). Perhaps this difference is due to the novelty effect: If children have little or no experience with robots, they may attend more to the robot and become more motivated by it, and thus learn more than when they become more familiar with robots (see Leite et al., 2013, for an overview of long-term interactions with robots). Multiple-session studies are thus required to rule out a short-lived novelty effect as a main cause of children's word learning in RALL studies.

Moreover, and crucial to this paper, there is evidence that RALL may be only effective for a subgroup of children (such as children who are motivated to play with the robot; Kanda et al., 2004), suggesting that individual characteristics of children may moderate the effects of RALL. It is possible that robots are useful language-education tools for certain children only, for example, depending on children's prior language knowledge and general (language) learning abilities. However, studies on the role of individual child characteristics in RALL, enabling the identification of such specific groups, are scarce. Most studies on adaptive learning focus on learners' age, gender, or cognitive or affective state during the learning task (e.g., Gordon et al., 2016; Ahmad et al., 2019), and not so much on learners' prior skills.

The current study, therefore, aims to add to the evidence regarding the effectiveness of robots in L2 teaching of young children in a vocabulary training spanning multiple sessions, by specifically focusing on the role of individual differences across children in skills related to the task at hand. We focused on three skills suggested by the literature to play an important part in language learning: children's first language (L1) vocabulary knowledge (Wolter, 2006), phonological short-term memory capacity (Gathercole and Baddeley, 1990; Service, 1992; Baddeley et al., 1998; Masoura and Gathercole, 2005; Gathercole, 2006; Verhagen et al., 2019), and selective attention (Schmidt, 1990; Robinson, 1995). We will examine whether these skills moderate any effects of RALL on children's learning of L2 words.

The current study follows up on a previous study using the same data (Vogt et al., 2019). In this previous study, the added value of a social robot and its iconic gestures for L2 vocabulary learning were investigated. Native Dutch-speaking five-year-old children were taught L2 English vocabulary in the domains of mathematical and spatial language in a series of seven short, individually administered lessons. Children were taught words through language games on a tablet in one of three conditions: 1) by themselves (the tablet-only condition); 2) with a robot that used deictic (pointing) gestures (the no-iconic-gestures condition); or 3) with a robot that used both deictic and iconic gestures (i.e., gestures depicting the target word; the iconic-gestures condition). In addition, a control group of children was included who did not receive the vocabulary training but played dancing games with the robot instead. Children in the experimental conditions were found to outperform children in the control condition on word-knowledge tasks in two post-tests, both directly after the training and two to four weeks later. However, there were no differences in word knowledge between children across the three experimental conditions on either one of these post-tests. Thus, no overall benefit of the robot's presence or its iconic gestures was found in Vogt et al. (2019).

In the present study, we extend this earlier study by Vogt et al. (2019) by investigating whether any effects of the robot's presence or its gestures would be moderated by children's (language) learning skills. Both the general research question on the added value of the robot and its iconic gestures (answered in Vogt et al., 2019) and the follow-up exploratory question on individual differences (answered in the current paper) were preregistered on AsPredicted[1]. As noted above, we considered three aspects of children's (language) learning skills, as moderator variables: L1 vocabulary knowledge, phonological memory, and selective attention. If effects of these variables are found, our findings will show the importance of taking into account individual differences in RALL and help tailor RALL to individual children to optimize learning outcomes. Below, we first describe how L1 word knowledge, phonological memory, and selective attention may play a role in L2 word learning, before we turn to our research question and hypotheses on how they may play a role in RALL in particular.

Learning an L2 is dependent on both the quality and quantity of the L2 input (Hoff, 2013; Unsworth, 2016) and on characteristics of the learner (i.e., the learner's cognitive and personality resources; Cummins, 1991). Prior L1 knowledge may help in L2 learning, as learners can map new L2 labels onto underlying concepts which they already acquired in their L1, provided that concepts are similar (Wolter, 2006). Besides conceptual similarity, similarity in word form between L1 and L2 can also aid in L2 learning, at least when this similarity also entails similarity in meaning (Brenders et al., 2011; Hemsley et al., 2013; Sheng et al., 2016). On the basis of these findings, children with larger L1 vocabularies are expected to learn more words from L2 vocabulary interventions than children with smaller L1 vocabularies. Children with larger L1 vocabularies can use their richer lexical and conceptual networks to disambiguate new input and to integrate it in existing knowledge. This phenomenon, found in particular for reading instruction but also in vocabulary learning (e.g., Penno et al., 2002), has been referred to as the Matthew effect (Stanovich, 2009).

Another factor relevant for L2 learning is children's phonological memory, defined as the capability to construct a phonological representation of speech sound sequences and to temporarily hold this representation active in memory for further processing (Gathercole and Baddeley, 1990; for a review on the relationship between phonological memory and word learning, see Gathercole, 2006). Phonological memory has been found to predict both L1 and L2 vocabulary learning (Gathercole and Baddeley, 1990; Service, 1992; Baddeley et al., 1998; Masoura and Gathercole, 2005; Gathercole, 2006; Verhagen et al., 2019). Phonological memory may aid L2 vocabulary learning, either directly or through its effect on L1 vocabulary knowledge, in particular if the learner is a novice and still has limited L2 vocabulary knowledge (Cheung, 1996; Masoura and Gathercole, 2005). Learners with substantial L2 vocabulary knowledge can rely on semantic, conceptual, or phonological similarities between novel words and words they have already learned, while novice learners cannot do this and thus have to rely more on their phonological memory (Masoura and Gathercole, 2005).

Finally, language learning in both L1 and L2 may depend on general learning abilities, in particular selective attention – a skill that has been considered the core of executive functions and working memory by some researchers (Garon et al., 2008; Mulder et al., 2014; Hendry et al., 2016; Cowan, 2017). Selective attention, defined as a domain-general, effortful mechanism of perceptual focusing, helps individuals to filter relevant information from irrelevant information in the encoding stage of linguistic information processing and supports processing in working memory. Language learning is thought to depend in part on automatic implicit processes (e.g., statistical learning), but attention can strengthen implicit learning (e.g., Lewkowicz and Hansen-Tift, 2012; Stevens and Bavelier, 2012), and learning may also depend on explicit processes that require attentional effort, especially in L2 learning at a later stage (e.g., the Noticing Hypothesis; Schmidt, 1990).

In the present study, we investigated whether individual differences in L1 word knowledge, phonological memory, and selective attention moderated the extent to which children benefited from the robot's presence and its iconic gestures during robot-assisted L2 learning. We used the data from Vogt et al. (2019), from all three experimental conditions (the tablet-only, no-iconic-gestures, and iconic-gestures condition), and the control condition. The choice to include a tablet-only condition was motivated by the fact that we had to work around limitations of the robot with regard to speech and object recognition (Kennedy et al., 2017; Wallbridge et al., 2017), which could only be resolved by including a tablet as an additional device for communication and interaction, as is explained more extensively in Vogt et al. (2019). Based on the findings in language learning research, discussed above, we expected that children with larger L1 vocabulary knowledge, larger phonological memory capacity, and a higher level of selective attention would learn more English words across all experimental conditions (i.e., conditions involving a robot and/or a tablet) than children scoring lower on these skills. We did not expect to see effects of L1 vocabulary knowledge, phonological memory, and selective attention for children in the control condition on their English word knowledge, as these children were not taught any English words. We contrasted the experimental conditions with the control condition to make sure that, if any moderator effects were found in the experimental conditions, they would pertain to the learning process, and not to the test taking.

In the remainder or this section, we will discuss our hypotheses regarding possible moderator effects in the robot-assisted vs. tablet-only conditions, before discussing our hypotheses regarding moderator effects in the iconic-gestures vs. no-iconic-gestures conditions. All our hypotheses are quite general, as our study is, to the best of our knowledge, the first to investigate the moderating effects of children's (language) learning skills on RALL. Our expectation was that the robot conditions offered children a more naturalistic and supportive language learning setting than the tablet-only condition (i.e., a setting in which the learner interacts with another being and which is grounded in the physical environment; Barsalou, 2008; Ellis, 1999; Gallaway and Richard, 1994; Hockema and Smith,

[1]https://aspredicted.org/6k93k.pdf

TABLE 1 | Background characteristics of the children in the four conditions.

	Tablet only	No iconic gestures	Iconic gestures	Control
N	53	54	54	32
n girls	29	26	22	18
M age (*SD*) in months	69.1 (4.4)	68.5 (4.7)	68.4 (4.8)	66.9 (4.7)
Age range in months	61–79	59–79	60–81	59–79
M standardized PPVT score	105.1 (12.3)	108.6 (11.7)	107.9 (14.5)	108.9 (14.0)
Parental education				
Academic level	60%	72%	74%	66%
Vocational level	33%	26%	20%	24%
Secondary school	7%	2%	6%	10%

Note. *Information on parental education of both parents was gathered through a questionnaire with a response rate of 65.8%, thus for 127 out of 193 children (*n *iconic-gestures condition = 40,* n *no-iconic-gestures condition = 32,* n *tablet-only condition = 34,* n *control condition = 21).*

2009; Iverson, 2010; Wellsby and Pexman, 2014), as the robot had a social presence and provided visual input (i.e., iconic and/or deictic gestures) in addition to the tablet. Our hypothesis, therefore, was that the presence of the robot would particularly benefit children who are poorer at language learning, that is, children with smaller L1 vocabulary knowledge, smaller phonological memory capacity, and a lower level of selective attention. These children in particular would need support in relating the novel (L2) words to their existing (L1) knowledge. Thus, we expected these children to show larger differences in learning outcomes between the robot-assisted conditions and the tablet-only condition compared to children higher in (language) learning abilities.

Our hypothesis with respect to the difference between the two robot-assisted conditions (i.e., the added value of the robot's iconic gestures) was that the iconic gestures would further add to the naturalistic language learning environment and its visual support, and therefore, would particularly benefit children poorer at language learning. Iconic gestures visualize words and help learners to relate novel words to existing concepts (Tellier, 2008; Macedonia et al., 2011; Rowe et al., 2013), which may benefit children poorer in language learning in particular (van Berkel-van Hoof et al., 2019). Thus, we expected children low in (language) learning abilities to show a larger difference in learning outcomes between the iconic-gestures condition and the no-iconic-gestures condition, compared to children with higher learning abilities.

METHODS

Participants

One hundred and ninety-three[2] monolingual Dutch preschoolers (95 girls) with an average age of 68.4 months (range 59–81 months, *SD* = 4.7 months) participated in the study. They were recruited from nine different schools in the Netherlands and were randomly assigned within schools to one of the four conditions, while ensuring a similar gender distribution over conditions. None of the schools taught English to preschool children. Parents indicated in a background questionnaire that most children received limited English input. Most children heard, with a maximum of 2 days per week, some English in the media or when parents used stand-alone words and phrases like "let's go". **Table 1** displays the background characteristics of the children divided over the four conditions. There were no significant differences between conditions in parental education, age, and gender, all $ps > 0.303$. Eleven additional children started the lessons but did not complete them due to illness, technological problems, or because they did not want to participate anymore (*n* iconic-gesture condition = 6, *n* no-iconic-gesture condition = 3, *n* tablet-only condition = 2). Three additional children were pre-tested but excluded from the experiment because they knew more than half of the target words at the pre-test. Informed consent for all children was obtained from parents/caretakers prior to data collection. The L2TOR project, in which this study was embedded, received ethical approval from Utrecht University's Ethics Committee under protocol number FETC16-039.

Overview of Experimental Sessions

The experiment consisted of a pre-test, seven tutoring sessions, and two post-tests (see **Figure 1** for an overview of the experiment). All children participated in a groupwise introduction prior to the pre-test (see Procedure for more information). In the pre-test, we measured children's knowledge of the L2 target words and their L1 vocabulary knowledge, phonological memory, and selective attention. The training was administered in one of four conditions, and had a between-subject design. In the experimental conditions, children played language games on a Microsoft Surface tablet: 1) by themselves; 2) with a robot that used deictic gestures (see the Robot section for more information on the robot used); or 3) with a robot that used iconic and deictic gestures. Children received on average two lessons per week over a period of on average 24 days (*SD* = 5.5 days). Children in a fourth, control condition did not play language games but danced with the robot during three sessions, once every week. Children's immediate learning outcomes were measured in a game concluding each lesson (which are beyond the scope of

[2]One child from the Vogt et al. (2019) data was excluded from the analyses in the current paper, as this child had a different home language in addition to Dutch, which may interfere with the (language) learning variables studied here.

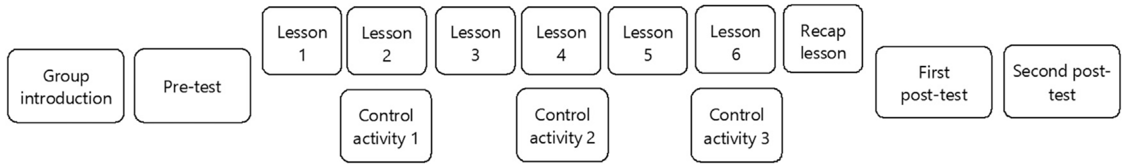


FIGURE 1 | Overview of the experiment.

TABLE 2 | Overview of the lesson series and target words.

Lesson	Location	Target words
One	Zoo	One, two, three, add, more, most
Two	Bakery	Four, five, take away, fewer, fewest
Three	Zoo	Big, small, heavy, light, high, low
Four	Fruit shop	On, above, below, next to, fall
Five	Forest	In front of, behind, walk, run, jump, fly
Six	Playground	Left, right, catch, throw, slide, climb
Seven	Photo book	Repetition of all target words

the current paper), and in a post-test one or two days after the seven tutoring lessons and a second post-test two to four weeks after the first post-test (M = 18.9 days, SD = 3.6 days) to measure retention over a longer period.

L2 Vocabulary Lessons

The lesson series consisted of seven individual lessons: six lessons in which new L2 vocabulary was provided, and one recap lesson in which all target words were repeated. Five or six target words were taught within each lesson, resulting in a total of 34 target words. The target words were chosen such that they were part of early mathematical and spatial language. This type of language – academic language – is highly important for later academic success (Hoff, 2013; Leseman et al., 2019). The overall theme of the lesson series was an area to be explored, with different locations for each lesson, such as "the zoo", "the bakery", etcetera. The locations were chosen such that they were familiar and relevant to young children. See **Table 2** for an overview of the lesson series, the locations, and the target words.

Each lesson consisted of four parts. First, the child was greeted, a reference was made to the previous lesson, and the location of the current lesson was introduced. Then, the new target words were modelled. New target words were first introduced by a pre-recorded speech sample of a native (Canadian) English speaker. The child was asked to repeat the target word, as this benefits productive recall of target words (Ellis and Beaton, 1993). Then, the child was instructed to perform several tasks on the tablet to practice the target words, for example, during the first lesson in the zoo, children had to put two elephants in a cage to practice the word "two". The tasks allowing children to practice the target words differed per target word. Some target words required manipulations on the tablet, while others allowed for more physical activity. For example, children were asked to act out running when being taught the word "running". The lessons concluded with a short test, to measure immediate learning outcomes. We will not discuss these immediate tests in this paper, as this would make our–already extensive–data set too complicated and would distract from the overall picture in which we were interested, namely the overall effect of (language) learning skills on robot-assisted word learning and retention, rather than immediate learning gains.

Each target word was repeated ten times throughout the lesson: nine times by the robot, and once by the native English speaker when it was introduced. Each target word reoccurred once in the following lesson and twice in the recap lesson. During the recap lesson, a photo book appeared on the tablet, which showed print screens from the previous lessons. Children had to practice repeating the target words once more during this recap lesson.

Robot

The robot used in this study was a Softbank Robotics NAO robot[3]. The robot was sitting in crouch position during the lesson series in a 90° angle to the right of the child, which was sitting on the floor facing the tablet that was positioned on an elevated surface.

The robot's responses had been preprogrammed, such that its responses and behaviors were consistent for all children. The robot was nearly autonomous; it behaved by responding to the child's actions on the tablet. The only function controlled by the experimenter was voice detection, as automatic speech recognition systems do not work reliably for children (Kennedy et al., 2017). This function was only used when children were asked to repeat the target words. The experimenter indicated, using a graphical user interface on a laptop computer, whether the child had produced sounds or not. The laptop computer was not in direct sight of the child (see **Figure 2**). The robot was introduced as Robin (which is a gender-neutral name in Dutch), being a peer that was going to learn English words together with the children.

The robot acted as a slightly more knowledgeable peer who understood the game usually faster than the child. As such, the robot performed several behaviors during the training: 1) talking to the child and explaining the tasks of the lesson; 2) pronouncing the target words; 3) providing feedback on the actions of the child; 4) pointing to the tablet while explaining what to do; 5) performing required manipulations in case the child failed to

[3]https://www.softbankrobotics.com/emea/en/nao

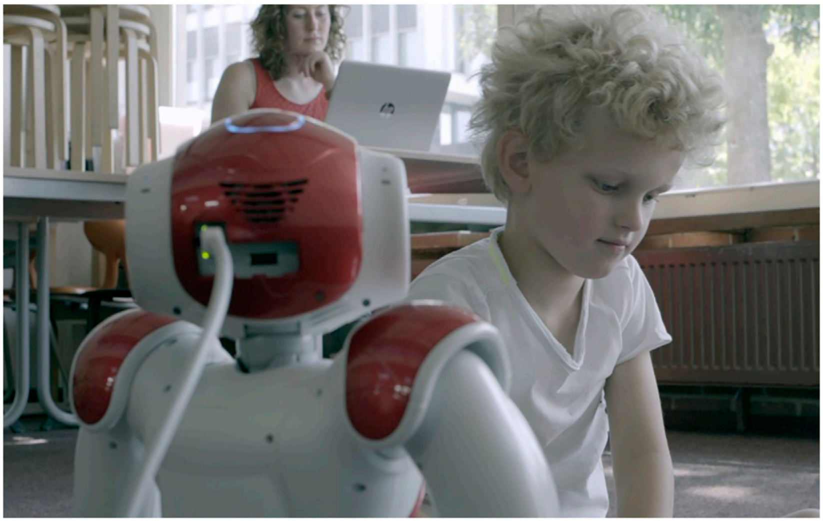


FIGURE 2 | A child engaging in a lesson with the robot.

perform a specific task. In case of the latter, the robot moved its arm above the tablet and any required manipulations "magically" occurred. In the iconic-gestures condition, the robot made an iconic gesture each time it pronounced a target word in English. These gestures were modeled after the gestures adults made in a gesture-elicitation procedure when they were asked to make an iconic gesture for each target word[4].

To ensure that the content and structure of the lessons were the same between the different conditions, in the tablet-only condition, the robot's voice was redirected through the tablet's speakers, and the robot itself was hidden from sight. Thus, the robot was used "behind the scenes" to operate the system, but children only saw and interacted with the tablet. In the robot-assisted conditions children thus interacted with the robot and the tablet, whereas in the tablet-only condition they interacted with the tablet only.

Measures

Pre-test Translation Task

To measure whether children knew the L2 English target words prior to the lesson series, we administered a translation task. In this task, children heard the 34 English target words one by one and were asked to translate them to Dutch. The target words were pre-recorded by a native speaker (different from the native speaker whose voice introduced the target words for the first time through the tablet) and played through a laptop computer. Two versions of this task were used, differing in word order. The first list of words was created by listing the target words randomly, and a second list was created by reversing the first list. Children were awarded one point per correct answer, yielding a maximum score of 34 points. Cronbach's alpha showed that the internal consistency of the task was excellent, $\alpha = 0.96$.

Post-test Translation Tasks

To measure how many L2 English target words the children learned during the lesson series, we administered two translation tasks: one from English to Dutch and one from Dutch to English. The task was the same as the pre-test translation task, except that children now also had to translate the words from Dutch to English. We did not include the Dutch-to-English translation task in the pre-test, because this would make the pre-test too long and difficult for the children (they were expected to know very few English words, and translating them from English to Dutch was expected to be a sufficient measure to assess their existing knowledge of the English target words). Both tasks were administered twice after the lesson series had ended, once during the first post-test and once during the second post-test. Children were awarded one point per correct answer, resulting in a maximum score of 34 points per task. Cronbach's alpha showed that the consistency of both tasks was excellent, $\alpha = 0.94$ at the first post-test and $\alpha = 0.95$ at the second post-test for the English-to-Dutch translation task, and $\alpha = 0.97$ at the first post-test and $\alpha = 0.98$ at the second post-test for the Dutch-to-English translation task.

Post-test Comprehension Task

We administered a comprehension task to measure children's receptive knowledge of the target words taught. The comprehension task was a picture-selection task in which we presented children with three images (still photos for most words, or short films in the case of verbs) on a laptop screen. Children then had to select the image corresponding to the target word they heard. Again, pre-recorded speech was used. A bilingual native English-Dutch speaker pronounced a Dutch carrier sentence "waar zie je" ("where do you see") followed by the

[4]Video recordings of the gestures made by the robot can be found at https://tiu.nu/l2tor-gestures.

target word in English. There were three trials for each target word, with different distractors each time. We selected half of the target words for this task to reduce children's fatigue, as a comprehension task consisting of all items would have been too long for the children. The target words included in the tests were chosen such that words from each lesson were included and that different types of words (verbs, adjectives, prepositions) were included. Two versions of this task were used, differing in word order: The first list of words was created by listing the target words randomly, and a second list was created by reversing the first list. Children were awarded one point per correct answer, resulting in a maximum score of 54 points. This task was administered during the first and second post-test. Cronbach's alpha showed that the consistency of the task was good, $\alpha = 0.84$ at the first post-test and $\alpha = 0.87$ at the second post-test.

L1 Vocabulary

We used the Dutch version of the Peabody Picture Vocabulary Test (PPVT-III-NL, Dunn et al., 2005) to measure children's Dutch receptive vocabulary knowledge. In this task, children are presented with four pictures and asked to select the picture corresponding to a word said by the experimenter. The task contains a total of seventeen sets, with each set consisting of twelve items. The test is adaptive, such that the starting set is chosen depending on the age of the child, and testing is stopped when the child makes nine or more errors within one set. The test is age-normed, with a mean of 100 and a standard deviation of 15. Cronbach's alpha is described in the test manual to be between 0.92 and 0.94. We used standardized scores in our analyses.

Phonological Memory

The Cross-Linguistic Nonword Repetition Task (CL-NWR) was used to measure phonological memory (Boerma et al., 2015; Chiat, 2015). The CL-NWR is a computerized task appropriate for young children, consisting of sixteen items, ranging from two to five syllables in length. Children hear a previously recorded, non-existing word via a laptop computer, and are asked to repeat it. Children receive two practice items (two one-syllable nonwords) before starting. Children's responses were scored online by the experimenter and they received one point for each word that they repeated correctly, yielding a maximum score of twelve. Cronbach's alpha showed that the consistency of this task was satisfactory, $\alpha = 0.76$. Ten percent of the data was scored independently by an additional researcher based on video recordings of the test. Inter-rater reliability was good with 89% agreement, $\kappa = 0.74$ [95% CI (0.663–0.819)], $p < 0.001$.

Selective Attention

A computerized visual search task was used to measure selective attention (Mulder et al., 2014). In this task, children were shown a display of animals on a laptop screen consisting of elephants, bears, and donkeys that were similar in color and size. Children were asked to find as many elephants as possible among distractor animals. Children were given three practice items and four test items that increased in difficulty. In the first two test items, 48 animals appeared on a six by eight grid. In the third item, 72 animals (similar in size to the first two test items) appeared on a nine by eight grid. In the last item, 204 animals (smaller in size than in the other three test items) appeared on a 12 by 17 grid. There were eight targets (elephants) in total in each test item. Each test item lasted 40 s. The experimenter encouraged children to search as quickly as possible and gave feedback according to a strict protocol. Elephants that were found were crossed off with a line by the experimenter. The number of targets located correctly per item was calculated and averaged across items, resulting in a maximum score of eight. Cronbach's alpha showed that the consistency of this task was good, $\alpha = 0.86$.

Procedure

Group Introduction of Robot

Prior to the individual sessions, the robot was introduced to all children in a group session. The robot introduced itself and did a dance with the children. The groupwise introduction served to familiarize children with the robot, and reduce potential anxiety during the individual sessions.

Pre-test

All children were tested individually by a trained experimenter in a quiet room in their schools. Children were administered the tasks in the following order: PPVT, pre-test translation task, selective-attention task, and CL-NWR. Furthermore, a perception questionnaire was administered (also during the first post-test) which measured the degree to which children anthropomorphized the robot. This questionnaire is beyond the scope of the current paper as it did not measure language skills or learning outcomes, and the results of this questionnaire can be found in Van den Berghe et al. (2020). The pre-test session lasted 30–40 min. Children got a sticker in reward for each task.

L2 Vocabulary Lessons

Each lesson was administered individually in a quiet room at the children's schools. At the start of the first session, the experimenter explained how the child could perform the requested actions on the tablet during the lessons (e.g., swiping and tapping), and helped the child to play the game. The experimenter was always present during the lessons to help children if needed, and to control the robot. The lesson could be paused if children needed a break. Each lesson lasted 15–20 min.

Control Activities

Children in the control condition participated in a total of three activities with the robot, each administered individually in a quiet room in the children's schools. In each session, the robot greeted the children, did a dance together with the child, and said goodbye. Each session lasted around five to 10 min.

First and Second Post-test

Children were administered the various tasks in the following order: the English-to-Dutch translation task, the Dutch-to-English translation task, and the comprehension task. During the first post-test the anthropomorphism questionnaire was also administered. Each session lasted around 30 min. Children got a sticker in reward for each task completed.

Analyses

We ran a MANOVA to compare the four groups of children on L1 vocabulary knowledge, phonological memory, selective attention, and pre-test scores. Children's scores on the comprehension task were compared against chance level (33%) using one-sample *t*-tests. To investigate differences in learning outcomes between the four conditions, we ran mixed-effect logistic regression models in the statistical package R (R Core Team, 2017) using the lme4 package (Bates et al., 2015). Dependent variables were children's binary (correct/incorrect) scores on the translation tasks and the comprehension task. The analyses were run separately for the translation tasks and the comprehension task, as they were assumed to measure different types of vocabulary knowledge. For both types of tasks, both assessments (the first and second post-test) were included.

Linear mixed-effects models included both fixed and random factors. The fixed-effect factors that were included in the models for the comprehension and translation tasks were condition (control, tablet-only, no-iconic-gestures, and iconic-gestures) and time (first and second post-test), with an interaction between them. For the translation task, target language (from English as source to Dutch as target, and vice versa) was included as an additional fixed-effect factor. The models were run separately for each of the three moderator variables (L1 vocabulary knowledge, phonological memory, and selective attention), as models with more than one moderator variable did not converge. The moderator variables were included as a fixed-effect factor in interaction with condition.

We included random factors and slopes by estimating a series of models with various combinations of random factors and slopes. We compared models by performing likelihood ratio tests that compared the goodness of fit using the ANOVA function in the base package (R Core Team, 2017). First, models were selected by checking whether the *p*-value from the likelihood ratio test was significant. Then, AIC and BIC values were compared, and the model with the smallest values were chosen. For the translation tasks, "participants", "target words", and "test item number" were included as random factors, and random slopes for target words (condition*target word). For the comprehension task, "participants", "target words", and "test item number" were included as random factors, and no random slopes were included as models including random slopes did not converge. We kept our models maximal, that is, we chose the models with the maximal random effects structure that converged (Barr et al., 2013).

We applied orthogonal sum-to-zero contrast coding to our categorical effects (i.e., condition, time, language; Schad et al., 2020), and all continuous variables (i.e., vocabulary knowledge, phonological memory, selective attention) were centered around zero (Baguley, 2012, pp. 590–621). For time, the first post-test (coded as −0.5) was contrasted with the second post-test (coded as 0.5). For condition, there were three contrasts: Contrast one contrasted the three experimental conditions (each coded as 0.25) with the control condition (coded as −0.75); Contrast two contrasted the two robot-assisted conditions (each coded as −0.33) with the tablet-only condition (coded as 0.66); and Contrast three contrasted the iconic-gestures condition (coded as −0.5) with the no-iconic-gestures condition (coded as 0.5). The number of iterations was increased to 100,000 using the bobyqa optimizer to solve issues of non-convergence (Powell, 2009).

The full results of each model can be found in the **Supplementary Tables S2–7**)[5]. The "ß" is an indicator of the effect size. To reduce the risk of Type-1 error when conducting multiple comparisons, we applied the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) with a false-discovery rate of 5%. The outcomes of this procedure can be found in the **Supplementary Tables S8–13**), and the calculations can be found online with the dataset[6].

RESULTS

Descriptive Analyses

Table 3 displays the descriptive statistics for all the variables included in the analyses for the children in each condition separately. A MANOVA showed no statistically significant differences in L1 vocabulary, phonological memory, selective attention, and English vocabulary pre-test scores between the conditions, F (12, 397) = 1.75, p = 0.054, η_p^2 = 0.05. For all conditions, children scored above chance level on the comprehension task on both the first and second post-test, all *p*s < 0.001, range *d*s = 1.49–2.83.

Effects of (Language) Learning Skills

First, we discuss the moderator effects of L1 vocabulary, phonological memory, and selective attention on the comparison of the experimental conditions versus the control condition. Then, we will discuss the moderator effects on the comparisons of the robot-assisted versus tablet-only conditions, and on the iconic-gestures versus no-iconic-gestures conditions. All outcomes can be found in the Appendix and the interactions are displayed in **Figure 3**.

Experimental Conditions vs. Control Condition

First, we investigated whether children's (language) learning skills moderated the effect of the vocabulary intervention itself. We expected an effect of children's (language) learning skills in the experimental conditions but not in the control condition, as only children in the experimental conditions received an L2 vocabulary training in which they could benefit from these skills. The models of the translation tasks showed statistically significant interactions between the moderator variables and condition. There were positive

[5]We also ran our model without moderator variables, and confirmed the results from Vogt et al. (2019; see **Supplementary Table S1** for the results).
[6]https://doi.org/10.17605/OSF.IO/GSNEK.

TABLE 3 | Means (standard deviation) on all the tasks in the pre-test and post-tests for the four conditions.

		Iconic	No iconic	Tablet-only	Control
Pre-test	L1 vocabulary	108.13 (12.54)	108.67 (11.83)	105.77 (11.92)	108.88 (13.96)
	Phonological memory	10.08 (2.97)	11.33 (2.86)	11.08 (2.13)	10.16 (3.22)
	Selective attention	6.48 (0.65)	6.82 (0.58)	6.67 (0.64)	6.61 (0.82)
	Translation En-Du	3.41 (3.05)	3.59 (3.14)	3.98 (2.74)	2.81 (2.83)
First post-test	Translation En-Du	7.54 (5.14)	7.83 (4.94)	7.91 (4.63)	3.81 (3.21)
	Translation Du-En	6.09 (4.15)	6.54 (4.28)	6.64 (4.01)	3.16 (2.27)
	Comprehension	29.39 (5.78)	29.50 (6.13)	29.53 (6.40)	25.03 (6.66)
Second post-test	Translation En-Du	8.20 (4.98)	8.02 (4.92)	8.57 (4.61)	4.34 (3.22)
	Translation Du-En	6.57 (4.60)	6.44 (4.59)	6.75 (4.22)	3.47 (2.13)
	Comprehension	30.54 (6.26)	29.69 (6.61)	30.30 (6.55)	26.00 (6.04)

Note. *The L1-vocabulary test is age-normed, with a mean of 100 and a standard deviation of 15. The maximum scores were 16 for the phonological-memory test, eight for the selective-attention test, 34 for each translation task, and 54 for the comprehension task (chance level for the latter task was 18).*

main effects for L1 vocabulary, $ß = 72.47$, $SE = 8.99$, $z = 8.06$, $p < 0.001$, phonological memory, $ß = 22.13$, $SE = 8.07$, $z = 2.74$, $p = 0.006$, and selective attention, $ß = 36.46$, $SE = 6.47$, $z = 5.63$, $p < 0.001$, but only for children in the experimental conditions, and not for those in the control condition, as expected. Note that the effects were only found for the translation tasks and not for the comprehension task.

Robot-Assisted Conditions vs. Tablet-Only Condition

Next, we investigated whether children's (language) learning skills moderated the effect of the robot's presence. We expected that the robot's presence would particularly benefit children with poorer skills. Differences in translation task scores were found between the robot-assisted and tablet-only conditions for two of the three moderators, that is, L1 vocabulary, $ß = -24.52$, $SE = 10.94$, $z = -2.24$, $p = 0.025$, and phonological memory, $ß = 26.72$, $SE = 10.53$, $z = 2.54$, $p = 0.011$. Children with smaller L1 vocabularies knew more words in the tablet-only condition than in the robot-assisted conditions, while children with larger L1 vocabularies knew slightly more words in the robot-assisted conditions than in the tablet-only condition. The effect was opposite for phonological memory: Children with better phonological memory knew more words in the tablet-only condition than in the robot-assisted conditions, while children with poorer phonological memory knew more words in the robot-assisted conditions than in the tablet-only condition. This effect was in line with our expectation. No effects were found for the comprehension task or for selective attention, contrary to our expectation.

Iconic-Gestures Condition vs. No-Iconic-Gestures Condition

Last, we investigated whether children's (language) learning skills moderated the effect of the robot's gestures. We expected that the iconic gestures would benefit children with poorer skills most. Children with larger L1 vocabularies knew more target words in the no-iconic gestures condition, while children with smaller L1 vocabularies knew more words in the iconic-gestures condition. This was indicated by the models run on the translation tasks, $ß = 31.99$, $SE = 9.16$, $z = 3.49$, $p < 0.001$, and comprehension task, $ß = 19.56$, $SE = 6.43$, $z = 3.05$, $p = 0.002$. Similarly, children with better phonological memory knew more target words in the no-iconic-gestures condition, while children with poorer phonological memory knew more words in the iconic-gestures condition, but only for the translation tasks, $ß = 40.52$, $SE = 14.99$, $z = 2.70$, $p = 0.007$. Those effects were in line with our expectations. Selective attention showed an opposite pattern: Children with better selective attention showed higher performance in the condition in which the robot used iconic gestures than in the condition in which it did not, again only on the translation tasks, $ß = -41.25$, $SE = 8.75$, $z = -4.71$, $p < 0.001$.

DISCUSSION

The aim of our experiment was to investigate whether individual differences in L1 word knowledge, phonological memory, and selective attention moderated whether children benefit from the robot's presence or its iconic gestures during robot-assisted L2 learning. Children in the present study were taught L2 English vocabulary through seven lessons in the form of tablet games, which they played either: 1) by themselves (the tablet-only condition); 2) together with a robot that used deictic gestures (the no-iconic-gestures condition); or 3) together with a robot that used both deictic and iconic gestures (the iconic-gestures condition). Furthermore, the children in the experimental conditions were compared to 4) a control group of children who did not play language games but played dancing games with the robot instead. Several statistically significant moderator effects were found, both expected and unexpected. We first discuss the general moderator effects of the (language) learning skills in the experimental conditions, then the moderator effects in the two robot-assisted conditions vs. the tablet-only condition, and lastly, the moderator effects in the iconic-gestures vs. no-iconic-gestures conditions. For the discussion of the general research question on the added value of the robot and its iconic gestures, see Vogt et al. (2019).

Regarding the overall effectiveness of the experimental conditions involving word-learning lessons compared to the control condition without word learning, we found the expected moderator effects: Children scoring high on L1 language knowledge, phonological memory, or selective

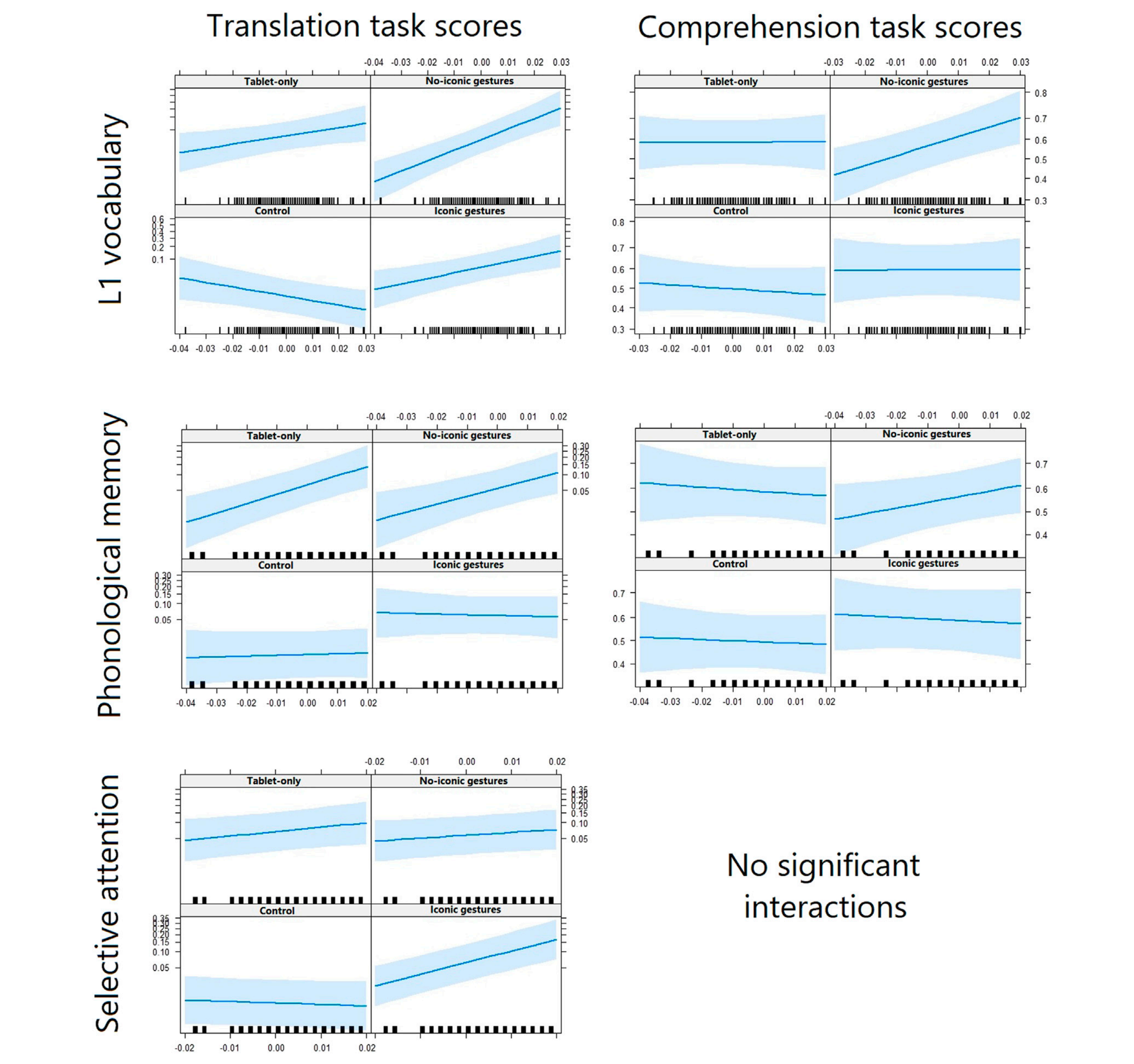


FIGURE 3 | Relations between children's English word-knowledge scores (*y*-axis) and (language) learning scores (*x*-axis), separated by condition and word-knowledge task.

attention, as assessed prior to the experiment, knew more words after the vocabulary lessons than children scoring low on these skills, in line with a vast body of literature that showed similar advantages in (second) language learning in general (Gathercole and Baddeley, 1990; Schmidt, 1990; Service, 1992; Robinson, 1995; Baddeley et al., 1998; Masoura and Gathercole, 2005; Gathercole, 2006; Wolter, 2006; Verhagen et al., 2019). No moderator effects were found in the control condition, which was expected because the control condition did not involve a word-learning intervention. The control condition, however, did involve an immediate and delayed post-test, similar to the experimental conditions. The lack of moderator effects in the control condition, therefore, supports the interpretation of the moderator effects in the experimental conditions as pertaining to the learning process, not to the test taking.

Regarding possible moderator effects between the three experimental conditions (i.e., the two robot-assisted conditions vs. the tablet-only condition), we expected the robot conditions to offer children a more naturalistic and supportive language-learning setting than the tablet-only condition (by grounding the interaction in the physical environment and allowing the learner to interact with another being; Barsalou, 2008; Ellis, 1999;

Gallaway and Richard, 1994; Hockema and Smith, 2009; Iverson, 2010; Wellsby and Pexman, 2014). We expected that this would particularly benefit children poorer at language learning (i.e., children with smaller L1 vocabulary knowledge, smaller phonological memory capacity, and a lower level of selective attention). The robot's presence particularly benefited children with larger L1 vocabularies or poorer phonological memory, while children with smaller L1 vocabularies or better phonological memory performed better in the tablet-only condition. These effects were only found for the translation tasks, and no effect was found for the comprehension tasks or for selective attention.

A possible explanation of why few effects were found is the prominent role of the tablet in our setup. It should be noted that the tablet was an essential device in the robot conditions, as technical limitations, in particular the lack of accurate speech perception (Kennedy et al., 2017) and object recognition for the type of robot we used (Wallbridge et al., 2017) required this extra device to enable interaction and communication. This may have limited the added value of the robot's social presence. We were aware that using a tablet was a risk that could limit the benefits of the robot. We could have chosen to teleoperate our robot using WoZ, allowing us to make a highly responsive, adaptive robot. However, in view of the educational relevance of the current study, we wanted to design a robot that could function nearly autonomously, such that it was more representative of the type of robots that can currently be implemented in schools. Many technological developments are still needed before a robot's full potential as an autonomous tutor in educational situations can be realized: Robots would need to be able to monitor the learner's speech, knowledge, mental state, emotions, and movements, and adapt their own behavior accordingly. In the meantime, a balance needs to be found between making robots as effective as possible, without losing their autonomy. Perhaps we need to change the design process. Rather than first focusing on what tasks would be ideal from an instruction perspective, we should consider earlier in the process what qualities the robot does and does not have, and design tasks that match these qualities optimally.

With respect to the two robot-assisted conditions (iconic gestures vs. no iconic gestures), we expected the iconic gestures to further add to the naturalistic language learning environment and its visual support, and therefore, to particularly benefit children poorer at language learning. Children with smaller L1 vocabularies or poorer phonological memory capacity as assessed prior to the experiment knew more English words in the iconic-gestures condition compared to the no-iconic-gestures condition, while children with larger L1 vocabularies or larger phonological memory capacity knew more English words in the no-iconic-gestures condition compared to the iconic-gestures condition. Note that these moderator effects were observed in addition to positive main effects of both conditions compared to the control condition, and suggest that iconic gesturing in RALL may support children with weaker (language) learning abilities. Thus, the iconic gestures particularly benefited children with poorer (language) learning abilities as expected. However, they disadvantaged children with stronger (language) learning abilities. Perhaps, the iconic gestures distracted these children, who did not need these gestures to learn from the learning task (similar to Kennedy et al., 2015). Anecdotal evidence supports this suggestion, as experimenters occasionally observed that children looked away when the robot was making its gestures. We are currently systematically investigating this by looking into children's engagement during the lessons (regarding both the learning task itself and the robot's involvement) and by conducting additional analyses to identify subgroups of children who possibly benefited from the iconic gestures (e.g., depending on their age).

Selective attention showed an opposite pattern. Children high in selective attention knew more English words than children low in selective attention in the iconic-gestures condition compared to the no-iconic-gestures condition. Note again that the moderator effect was found in addition to positive main effects of both experimental conditions relative to the control condition. A possible explanation points again to the distracting effect the iconic gestures may have had on children's word learning in this study (cf. Schmidt, 1990; Robinson, 1995). Children high in selective attention may have been better able to profit from the additional cues, which assumingly required attentional effort to perceive and interpret, and/or may have been less distracted by the extra information provided. Children low in selective attention may have been less capable in figuring out what the meaning was of the gestures and/or were more easily distracted by the gestures and the extra time it took the robot to perform these gestures. If true, this suggests that implementing iconic gestures benefits children with good attention skills, but disadvantages children with poorer attention skills. The results of benefiting some children while disadvantaging others highlights the importance of making adaptive robot-assisted lessons, which is in line with the conclusions of other recent studies that found limited benefits of a robot's gestures for language learning (de Wit et al., 2020; Demir-Lira et al., 2020).

The present study reveals moderator effects of children's (language) learning skills on the effectiveness of RALL. The findings, however, show a mixed pattern for the three (language) learning skills examined in this study. An open question is how children would respond to a robot (with or without iconic gestures) compared to a tablet or other non-robot condition if they have a mix of skills that are associated with opposite effects of robot-assisted instruction. For example, if a child both has a small L1 vocabulary and poor selective attention, it is unclear how they would respond to instruction by a robot with iconic gestures. On the one hand, the child could benefit from the iconic gestures. On the other hand, they would struggle to benefit from this additional, potentially distracting information. For future RALL research it is recommendable to identify profiles of skills in children and examine which profiles match best particular approaches to RALL. The overall results of the present study and Vogt et al. (2019) reveal that using robot tutoring in L2 learning programs for young children still has a long way to go. Designing the lesson series around the NAO robot, given the current state of technology, put severe constraints on the design of the lessons, required the use of a tablet for communication, and necessitated strong standardization. Traditional vocabulary training interventions may include more diverse activities that benefit learning and motivation, such as moving around and joint playing with objects. The robot was not yet capable of such activities in our study, and therefore, the only difference between the three experimental conditions was that children did not receive non-verbal support in the tablet-only condition through the robot's social presence and its (iconic and) deictic gestures. Overall, children did not benefit from the robot's presence, as was first reported in Vogt et al. (2019). The present study, however, reveals that children's (language) learning abilities may moderate the effects of the presence of a robot and its iconic gestures. Future studies on RALL will likely benefit from technological advancements that allow RALL to incorporate more elements of effective traditional vocabulary training interventions, and to adapt optimally to individual learners' skill profiles.

Our study is one of the first to investigate whether individual differences in children's (language) learning abilities moderate the added value of a robot and its iconic gestures for L2 vocabulary learning in a multiple sessions and well-powered experiment. Taken together, the results suggest that the study of individual differences and moderators is highly relevant, as they showed that children's (language) learning skills moderated the effect of the robot's presence and iconic gestures: Depending on their (language) learning skills, some children benefited from the robot's presence and iconic gestures, while some children appeared to be distracted by them. It is likely that the effects of the robot are different for different children and adaptation to children's learning profiles is warranted. Indeed, one of the real advantages of robots is that they can play different roles for different types of learners if programmed to do so. The present results should be replicated before any firm conclusions can be drawn. The study of individual differences is standard practice in educational sciences and developmental psychology, and could add to studies on the design of adaptive robots for educational practice.

DATA AVAILABILITY STATEMENT

The dataset and Benjamini-Hochberg calculations that support the findings of this study are openly available at OSF (https://doi.org/10.17605/OSF.IO/GSNEK). Further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Utrecht University's Ethics Committee (protocol number FETC16-039). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

RB, OO-P, JV, MH, JW, BW, PV, EK, and PL conceptualized the research; RB, OO-P, MH, JW, and BW collected the data, RB wrote the paper; SB analyzed the data; OO-P, JV, SB, MH, JW, BW, PV, EK, and PL critically reviewed the paper.

FUNDING

This project received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 688014.

ACKNOWLEDGMENTS

We would like to thank all members of the L2TOR team for their support. We are grateful to Laurette Gerts, Annabella Hermans, Esmee Kramer, Madee Kruijt, Marije Merckens, David Mogendorff, Sam Muntjewerf, Reinjet Oostdijk, Laura Pijpers, Chani Savelberg, Robin Sonders, Sirkka van Straalen, Sabine Verdult, Esmee Verheem, Pieter Wolfert, Hugo Zijlstra, and Michelle Zomers for their help in collecting and processing the data of this experiment. We would also like to thank all the schools, children, and parents of the children that participated in this experiment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2021.676248/full#supplementary-material

REFERENCES

Ahmad, M. I., Mubin, O., Shahid, S., and Orlando, J. (2019). Robot's Adaptive Emotional Feedback Sustains Children's Social Engagement and Promotes Their Vocabulary Learning: a Long-Term Child-Robot Interaction Study. *Adaptive Behav.* 27, 243–266. doi:10.1177/1059712319844182

Baddeley, A., Gathercole, S., and Papagno, C. (1998). The Phonological Loop as a Language Learning Device. *Psychol. Rev.* 105, 158–173. doi:10.1037/0033-295x.105.1.158

Baguley, T. S. (2012). *Serious Stats: A Guide to Advanced Statistics for the Behavioral Sciences*. Palgrave Macmillan. doi:10.1007/978-0-230-36355-7

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random Effects Structure for Confirmatory Hypothesis Testing: Keep it Maximal. *J. Mem. Lang.* 68, 255–278. doi:10.1016/j.jml.2012.11.001

Barsalou, L. W. (2008). Grounded Cognition. *Annu. Rev. Psychol.* 59, 617–645. doi:10.1146/annurev.psych.59.103006.093639

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using Lme4. *J. Stat. Softw.* 67, 1–48. doi:10.18637/jss.v067.i01

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social Robots for Education: A Review. *Sci. Robot.* 3, 5954. doi:10.1126/scirobotics.aat5954

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 (Methodological)* 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., and Blom, E. (2015). A Quasi-Universal Nonword Repetition Task as a Diagnostic Tool for Bilingual Children Learning Dutch as a Second Language. *J. Speech Lang. Hear. Res.* 58, 1747–1760. doi:10.1044/2015_jslhr-l-15-0058

Brenders, P., van Hell, J. G., and Dijkstra, T. (2011). Word Recognition in Child Second Language Learners: Evidence from Cognates and False Friends. *J. Exp. Child Psychol.* 109, 383–396. doi:10.1016/j.jecp.2011.03.012

Cheung, H. (1996). Nonword Span as a Unique Predictor of Second-Language Vocabulary Language. *Dev. Psychol.* 32, 867–873. doi:10.1037/0012-1649.32.5.867

Chiat, S. (2015). "Non-word Repetition," in *Methods for Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. Editors S. Armon-Lotem, J. de Jong, and N. Meir (Bristol: Multilingualism Matters), 227–250.

Cowan, N. (2017). The many Faces of Working Memory and Short-Term Storage. *Psychon. Bull. Rev.* 24, 1158–1170. doi:10.3758/s13423-016-1191-6

Cummins, J. (1991). "Interdependence of First- and Second-Language Proficiency in Bilingual Children," in *Interdependence of First-And Second-Language Proficiency in Bilingual Children. Language Processing in Bilingual Children*. Editor E. Bialystok (Cambridge University Press), 70–89. doi:10.1017/cbo9780511620652.006

de Wit, J., Brandse, A., Krahmer, E., and Vogt, P. (2020). Varied Human-like Gestures for Social Robots: Investigating the Effects on Children's Engagement and Language Learning. In Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction. Cambridge, UK. doi:10.1145/3319502.3374815

de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., et al. (2018). The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies. In Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction. Chicago, IL. doi:10.1145/3171221.3171277

Demir-Lira, Ö. E., Kanero, J., Oranç, C., Koşkulu, S., Franko, I., Göksun, T., et al. (2020). L2 Vocabulary Teaching by Social Robots: The Role of Gestures and On-Screen Cues as Scaffolds. *Front. Educ.* 5. doi:10.3389/feduc.2020.599636

Dunn, L. M., Dunn, L. M., and Schlichting, L. (2005). *Peabody Picture Vocabulary Test-III-NL*. Amsterdam: Pearson.

Ellis, N., and Beaton, A. (1993). Factors Affecting the Learning of Foreign Language Vocabulary: Imagery Keyword Mediators and Phonological Short-Term Memory. *The Q. J. Exp. Psychol. Section A* 46, 533–558. doi:10.1080/14640749308401062

Ellis, R. (1999). *Learning a Second Language through Interaction*. John Benjamins Publishing. doi:10.1075/sibil.17

Gallaway, C., and Richard, B. J. (1994). *Input and Interaction in Language Acquisition*. Cambridge University Press.

Garon, N., Bryson, S. E., and Smith, I. M. (2008). Executive Function in Preschoolers: a Review Using an Integrative Framework. *Psychol. Bull.* 134, 31–60. doi:10.1037/0033-2909.134.1.31

Gathercole, S. E., and Baddeley, A. D. (1990). The Role of Phonological Memory in Vocabulary Acquisition: A Study of Young Children Learning New Names. *Br. J. Psychol.* 81, 439–454. doi:10.1111/j.2044-8295.1990.tb02371.x

Gathercole, S. E. (2006). Nonword Repetition and Word Learning: The Nature of the Relationship. *Appl. Psycholinguistics* 27, 513–543. doi:10.1017/s0142716406060383

Gordon, G., Breazeal, C., and Engel, S. (2015). Can Children Catch Curiosity from a Social Robot?. in Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. Portland, OR. doi:10.1145/2696454.2696469

Gordon, G., Spaulding, S., Kory Westlund, J. M., Lee, J. J., Plummer, L., Martinez, M., et al. (2016). Affective Personalization of a Social Robot Tutor for Children's Second Language Skills. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, AZ.

Han, J.-H., Jo, M.-H., Jones, V., and Jo, J.-H. (2008). Comparative Study on the Educational Use of home Robots for Children. *J. Inf. Process. Syst.* 4, 159–168. doi:10.3745/jips.2008.4.4.159

Hemsley, G., Holm, A., and Dodd, B. (2013). Conceptual Distance and Word Learning: Patterns of Acquisition in Samoan-English Bilingual Children. *J. Child. Lang.* 40, 799–820. doi:10.1017/s0305000912000293

Hendry, A., Jones, E. J. H., and Charman, T. (2016). Executive Function in the First Three Years of Life: Precursors, Predictors and Patterns. *Dev. Rev.* 42, 1–33. doi:10.1016/j.dr.2016.06.005

Hockema, S. A., and Smith, L. B. (2009). Learning Your Language, Outside-In and Inside-Out. *Linguist* 47, 453–479. doi:10.1515/ling.2009.016

Hoff, E. (2013). Interpreting the Early Language Trajectories of Children from Low-SES and Language Minority Homes: Implications for Closing Achievement Gaps. *Dev. Psychol.* 49, 4–14. doi:10.1037/a0027238

Hyun, E., Kim, S., Jang, S., and Park, S. (2008). Comparative Study of Effects of Language Instruction Program Using Intelligence Robot and Multimedia on Linguistic Ability of Young Children. In Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication. Munich, Germany.

Iverson, J. M. (2010). Developing Language in a Developing Body: The Relationship between Motor Development and Language Development. *J. Child. Lang.* 37, 229–261. doi:10.1017/s0305000909990432

Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. (2004). Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Human-comp. Interaction* 19, 61–84. doi:10.1207/s15327051hci1901&2_4

Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., and Göksun, T. (2018). Social Robots for Early Language Learning: Current Evidence and Future Directions. *Child. Dev. Perspect.* 12, 146–151. doi:10.1111/cdep.12277

Kennedy, J., Baxter, P., and Belpaeme, T. (2015). The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Portland, OR. doi:10.1145/2696454.2696457

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., et al. (2017). Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. in Proceedings of the 2017 ACM. Vienna, Austria: IEEE Human Robot Interaction Conference. doi:10.1145/2909824.3020229

Kory Westlund, J., Dickens, L., Jeong, S., Harris, P., DeSteno, D., and Breazeal, C. (2015). A Comparison of Children Learning New Words from Robots, Tablets, and People. In Conference Proceedings of New Friends: The 1st International Conference on Social Robots in Therapy and Education. Almere, Netherlands.

Kory Westlund, J. M., Dickens, L., Jeong, S., Harris, P. L., DeSteno, D., and Breazeal, C. L. (2017). Children Use Non-verbal Cues to Learn New Words from Robots as Well as People. *Int. J. Child-Computer Interaction* 13, 1–9. doi:10.1016/j.ijcci.2017.04.001

Leite, I., Martinho, C., and Paiva, A. (2013). Social Robots for Long-Term Interaction: A Survey. *Int. J. Soc. Robotics* 5, 291–308. doi:10.1007/s12369-013-0178-y

Leseman, P. P. M., Henrichs, L. F., Blom, E., and Verhagen, J. (2019). "Young Monolingual and Bilingual Children's Exposure to Academic Language as Related to Language Development and School Achievement," in *Young Mono- and Bilingual Children's Exposure to Academic Language as Related to Language Development and School Achievement.* Learning Through Language. Editors V. Grøver, P. Ucelli, M. Rowe, and E. Lieven (Cambridge University Press), 205–217. doi:10.1017/9781316718537.019

Lewkowicz, D. J., and Hansen-Tift, A. M. (2012). Infants Deploy Selective Attention to the Mouth of a Talking Face when Learning Speech. *Proc. Natl. Acad. Sci.* 109, 1431–1436. doi:10.1073/pnas.1114783109

Leyzberg, D., Spaulding, S., Toneva, M., and Scassellati, B. (2012). *The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains*. Sapporo, Japan: Proceedings of the Annual Meeting of the Cognitive Science Society.

Macedonia, M., Müller, K., and Friederici, A. D. (2011). The Impact of Iconic Gestures on Foreign Language Word Learning and its Neural Substrate. *Hum. Brain Mapp.* 32, 982–998. doi:10.1002/hbm.21084

Marulis, L. M., and Neuman, S. B. (2010). The Effects of Vocabulary Intervention on Young Children's Word Learning. *Rev. Educ. Res.* 80, 300–335. doi:10.3102/0034654310377087

Masoura, E. V., and Gathercole, S. E. (2005). Contrasting Contributions of Phonological Short-Term Memory and Long-Term Knowledge to Vocabulary Learning in a Foreign Language. *Memory* 13, 422–429. doi:10.1080/09658210344000323

Mulder, H., Hoofs, H., Verhagen, J., van der Veen, I., and Leseman, P. P. M. (2014). Psychometric Properties and Convergent and Predictive Validity of an Executive Function Test Battery for Two-Year-Olds. *Front. Psychol.* 5. doi:10.3389/fpsyg.2014.00733

Penno, J. F., Wilkinson, I. A. G., and Moore, D. W. (2002). Vocabulary Acquisition from Teacher Explanation and Repeated Listening to Stories: Do They Overcome the Matthew Effect?. *J. Educ. Psychol.* 94, 23–33. doi:10.1037/0022-0663.94.1.23

Powell, M. J. (2009). *The BOBYQA Algorithm for Bound Constrained Optimization without Derivatives*. Cambridge: Technical Report, Department of Applied Mathematics and Theoretical Physics, University of Cambridge.

R Core team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.r-project.org/.

Robinson, P. (1995). Attention, Memory, and the "Noticing" Hypothesis. *Lang. Learn.* 45, 283–331. doi:10.1111/j.1467-1770.1995.tb00441.x

Rowe, M. L., Silverman, R. D., and Mullan, B. E. (2013). The Role of Pictures and Gestures as Nonverbal Aids in Preschoolers' Word Learning in a Novel Language. *Contemp. Educ. Psychol.* 38, 109–117. doi:10.1016/j.cedpsych.2012.12.001

Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R. (2020). How to Capitalize on A Priori Contrasts in Linear (Mixed) Models: A Tutorial. *J. Mem. Lang.* 110, 104038. doi:10.1016/j.jml.2019.104038

Schmidt, R. W. (1990). The Role of Consciousness in Second Language Learning1. *Appl. Linguistics* 11, 129–158. doi:10.1093/applin/11.2.129

Service, E. (1992). Phonology, Working Memory, and Foreign-Language Learning. *Q. J. Exp. Psychol. Section A* 45, 21–50. doi:10.1080/14640749208401314

Sheng, L., Lam, B. P. W., Cruz, D., and Fulton, A. (2016). A Robust Demonstration of the Cognate Facilitation Effect in First-Language and Second-Language Naming. *J. Exp. Child Psychol.* 141, 229–238. doi:10.1016/j.jecp.2015.09.007

Stanovich, K. E. (2009). Matthew Effects in reading: Some Consequences of Individual Differences in the Acquisition of Literacy. *J. Edu.* 189, 23–55. doi:10.1177/0022057409189001-204

Stevens, C., and Bavelier, D. (2012). The Role of Selective Attention on Academic Foundations: A Cognitive Neuroscience Perspective. *Dev. Cogn. Neurosci.* 2, S30–S48. doi:10.1016/j.dcn.2011.11.001

Tanaka, F., and Matsuzoe, S. (2012). Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning. *Jhri* 1, 78–95. doi:10.5898/jhri.1.1.tanaka

Tellier, M. (2008). The Effect of Gestures on Second Language Memorisation by Young Children. *Gest* 8, 219–235. doi:10.1075/gest.8.2.06tel

Unsworth, S. (2016). "Quantity and Quality of Language Input in Bilingual Language Development," in *Quantity and Quality of Language Input in Bilingual Language Development. Bilingualism across the Lifespan: Factors Moderating Language Proficiency*. Editors E. Nicoladis and S. Montanari (Washington, DC: De Gruyter Mouton), 103–121. doi:10.1037/14939-007

van Berkel-van Hoof, L., Hermans, D., Knoors, H., and Verhoeven, L. (2019). Effects of Signs on Word Learning by Children with Developmental Language Disorder. *J. Speech Lang. Hear. Res.* 62 (6), 1798–1812. doi:10.1044/2019_jslhr-l-18-0275

van den Berghe, R., de Haas, M., Oudgenoeg-Paz, O., Krahmer, E., Verhagen, J., Vogt, P., et al. (2020). A Toy or a Friend? Children's Anthropomorphic Beliefs about Robots and How These Relate to Second-language Word Learning. *J. Comput. Assist. Learn.* 37, 1–15. doi:10.1111/jcal.12497

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., and Leseman, P. (2019). Social Robots for Language Learning: A Review. *Rev. Educ. Res.* 89, 259–295. doi:10.3102/0034654318821286

Verhagen, J., Boom, J., Mulder, H., de Bree, E., and Leseman, P. (2019). Reciprocal Relationships between Nonword Repetition and Vocabulary during the Preschool Years. *Dev. Psychol.* 55, 1125–1137. doi:10.1037/dev0000702

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., et al. (2019). Second Language Tutoring Using Social Robots: a Large-Scale Study. In 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Daegu, South Korea. doi:10.1109/hri.2019.8673077

Wallbridge, C. D., Lemaignan, S., and Belpaeme, T. (2017). Qualitative Review of Object Recognition Techniques for Tabletop Manipulation. In Proceedings of the 5th International Conference on Human Agent Interaction. Bielefeld, Germany. doi:10.1145/3125739.3132593

Wellsby, M., and Pexman, P. M. (2014). Developing Embodied Cognition: Insights from Children's Concepts and Language Processing. *Front. Psychol.* 5. doi:10.3389/fpsyg.2014.00506

Wolter, B. (2006). Lexical Network Structures and L2 Vocabulary Acquisition: The Role of L1 Lexical/conceptual Knowledge. *Appl. Linguistics* 27, 741–747. doi:10.1093/applin/aml036

Zhexenova, Z., Amirova, A., Abdikarimova, M., Kudaibergenov, K., Baimakhan, N., Tleubayev, B., et al. (2020). A Comparison of Social Robot to Tablet and Teacher in a New Script Learning Context. *Front. Robotics AI* 7 (99). doi:10.3389/frobt.2020.00099

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 in Robotics and AI

ORIGINAL RESEARCH
published: 03 September 2021
doi: 10.3389/frobt.2021.680509

Robots for Foreign Language Learning: Speaking Style Influences Student Performance

Kerstin Fischer[1]*, *Oliver Niebuhr*[2] *and Maria Alm*[1]

[1]Department of Design and Communication, University of Southern Denmark, Sonderborg, Denmark, [2]Centre for Industrial Electronics, University of Southern Denmark, Sonderborg, Denmark

Much previous research suggests that teachers' individual characteristics may affect students' performance; however, which factors are particularly helpful is as yet unclear and methodologically very difficult to assess. In this paper, we study the effects of robots' speaking styles when instructing students on a task. 40 participants saw a brief video in which a robot presented its instructions either in a charismatic or a not so charismatic speaking style. Participants' task was then to produce foreign language sentences on the basis of visualizations of the prosodic properties of these sentences. A subsequent analysis of participants' productions shows that language learners' performance was significantly better when the robot had delivered its instructions in a charismatic voice. The results suggest not only that a charismatic speaking style may be crucial for teachers in general and hence one of the factors causing the interpersonal variation between teachers, but also that students can benefit from instructions by robots delivered in a charismatic speaking style.

Keywords: charisma, social robots, language learning, intonation, speaking style

OPEN ACCESS

Edited by:
Mohamed Chetouani, Université Pierre et Marie Curie, France

Reviewed by:
Khiet Phuong Truong, University of Twente, Netherlands
Yomna Abdelrahman, Munich University of the Federal Armed Forces, Germany

***Correspondence:**
Kerstin Fischer
kerstin@sdu.dk

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: *14 March 2021*
Accepted: *10 August 2021*
Published: *03 September 2021*

Citation:
Fischer K, Niebuhr O and Alm M (2021) Robots for Foreign Language Learning: Speaking Style Influences Student Performance. Front. Robot. AI 8:680509. doi: 10.3389/frobt.2021.680509

INTRODUCTION

With robots becoming more prevalent in our daily lives and possibly our classrooms (e.g., Leyzberg et al., 2012), the question to what extent robots can facilitate learning and teaching has gained some attention (cf. Belpaeme et al., 2018 for an overview). However, so far, the use of robots in pedagogical settings has not only yielded positive results; especially in the area of language learning, i.e., in learning to interact in a foreign language, previous studies have been rather discouraging. For instance, Rosenthal-von der Pütten et al. (2016) find no positive effect of using robots for language learning, and (Jiyoung and Jeonghye, 2015) find that because of the reliance on synthesized speech, robots are not useful to teach students native-like speaking competence.

That previous work has not only found robots to be successful teachers or tutors may thus be due to the way robots speak; speech is not only the prevalent medium of communication in the classroom and the target of the language learning process, but it also conveys a lot of information about the speaker, like, for instance, the speaker's gender, geographic origin, social class and speaker personality (Sutton et al., 2019). Of these, especially teacher personality has been shown to influence students' success (e.g., Lee et al., 2013). Research on teaching has long noted that individual teacher characteristics may play a role in pedagogy, and scholars have tried to identify what aspects of teacher personalities may have an effect on the learning process. Suspected factors include teacher's empathy, organization, adaptability, fostering of community, autonomy and enthusiasm (Klassen et al., 2018), and occasionally also charisma is noted as a potential factor (Towler et al., 2014; Lin and Huang 2017). That the way of speaking can have an effect has been suspected especially by experts on charismatic speech (e.g., Rosenberg and Hirschberg 2009).

However, there is no systematic investigation of the impact of the teacher's speaking styles on students' performance. There are at least three reasons for this: The first, methodological, reason is that very few speakers can manipulate their speaking styles to produce a charismatic speaking style in one situation and not in another, thus making the controlled investigation of speaking style difficult. The second reason is that for a long time, it has not been entirely clear what exactly makes a speaking style charismatic. The third reason is that a potential relationship between a charismatic speaking style and student performance does not matter if a charismatic speaking style cannot be taught to teachers, which is assumed by many (but see Abelin 2018).

In the current study, we circumvent the first problem by using robots as teachers. Robots, in contrast to people, can be manipulated at will, and they can produce identical output in comparable ways as often as necessary to all participants alike. Robots are therefore excellent tools for the study of effects of certain linguistic behaviors, such as speaking styles (cf. also Andrist et al., 2013, 2015; Fischer and Niebuhr 2020). In particular, robots' speech can be presynthesized and then manipulated by a prosody expert in order to match a particular speaking style. As for the second problem area, recent research has provided evidence for a short list of factors that contribute to charismatic speech (Niebuhr et al., 2016; Berger et al., 2017). In our own previous work, we demonstrated successfully that the speech features identified in the model make robot speech more persuasive (Fischer et al., 2020). Regarding the third problem area, robots can again provide the solution: Irrespective of whether a charismatic speaking style can or cannot be taught, if we have robots deliver the instructions, we only need to know how to generate charismatic robot utterances. By using robots to investigate the effects of teachers' speaking styles on student performance we furthermore shed light on how robots themselves can be employed to facilitate learning.

Consequently, in the current study, we focus on the characteristics of the instructions delivered by robots in a foreign language teaching context. In particular, we show that a more charismatic presentation of the task increases the correctness of students' performance significantly. In our study, the students' task is to interpret visualizations of the prosodic realization of questions in English, which generally constitutes a problem for them. A robot introduces the participants to the task, either using a charismatic speaking style or in a speaking style that is less charismatic. The results show that students who heard the charismatic robot produced significantly better results than students who heard the introduction by the less charismatic robot. Thus, the degree to which a teacher speaks charismatically can influence students' performance.

PREVIOUS WORK

Previous work concerns the role of teachers' charisma in language learning situations and especially on the effects of speaking styles on learning, as well as the roles of robots in language teaching.

Teacher Characteristics and Charismatic Speech

A review of previous work by Qardaku (2019) aims to narrow down what it means for teachers to be charismatic. Her analysis suggests that charismatic, and hence inspirational, teachers have to be experts, transmit enthusiasm for their topic and cultivate positive relationships with their students, reflect on their practices, and make learning meaningful to the students. None of these characteristics relates to the teacher's speaking style. Also other work on charisma in teaching leaves out the verbal dimension; for example, Nissim and Simon (2019) try to identify the characteristics that teachers and teacher-leaders have in common and find that charisma is expected of a teacher-leader, and certain personality characteristics are expected of good leaders and teacher-leaders to the same extent, but that the good teacher-leader should also be able to empower students - somewhat in line with Qardaku's (2019) suggestions. Similarly, Lin and Huang (2016) take it to be uncontroversial that a charismatic teacher is characterized by "knowledge, character, humour and teaching method," again without any reference to speaking style. Concerning these four characteristics, they find that they are positively related to interest in a subject. In a follow-up study, Lin and Huang (2017) find that the same features also influence students' attitude to calculus learning.

Towler et al. (2014) study the effects of what they consider charismatic content on the evaluation of the trainer and on recall and transfer immediately after the presentation and a week later. Their manipulations concern the articulation of a higher vision, positive emotional expression, emphasis of the importance of the contents, storytelling, use of metaphors, raising of expectations and encouraging of innovative thinking and the provision of encouragement and support. Students listened to a 15 min course on statistics software either with or without the charismatic features. The results show better evaluations of the charismatic teacher as well as better retention and transfer a week later. It is however unclear to what extent the trainer also used "appropriate vocal intonation," which the authors also consider to constitute a trait of charismatic trainers.

We can conclude that as yet, there is no systematic investigation of the relationship between speaking style and student performance, and thus we don't know the impact of a charismatic speaking style on teaching.

Robots as Tutors in Language Learning

Robots have been found to be generally engaging for students, and to have a positive influence even on their performance (e.g., Baxter et al., 2017; De Haas et al., 2020). However, robots can play various different roles in language learning situations. Belpaeme et al. (2018) provide an overview of current work on robots as language tutors (compared to as peers, as in Baxter et al.'s study) and conclude that the effects reported so far are rather small. Similarly, in a study in which the robot served as a teacher, Rosenthal-von der Pütten et al. (2016) find no advantage of an embodied robot over a computer simulation and in general neither application led students to improve their language skills. Also a large-scale study by Vogt et al. (2019) with 194 children yielded no advantages of a robot over a tablet, and the iconic gestures the robot was using (for instance, for "add", "behind" or "running") had no impact on children's word learning performance. Here the robot taught 5–6 year old children English words for already known concepts. In the control condition, children sang songs together with the experimenters and were not exposed to the English words at all. Another study by De Haas et al. (2020) finds that the variability of the feedback the robot provided in an animal name learning game had an effect on children's engagement in the game, but the feedback itself did not affect their learning gain, which was small but present in all conditions.

Rosenthal-von der Pütten et al. (2016) also investigated whether synthesized speech constitutes a problem for the language learners. They first synthesized the robot utterances, then had a human speaker speak them with similar speech characteristics in order to make the stimuli as comparable as possible. They find no differences in participants' self-reported experiences, on alignment with syntactic and lexical features of the robot's utterances and students' learning gains. Thus, the synthesized utterances performed no worse than the human utterances. In contrast, In and Han (2015) found the range in speech melody in synthesized speech to be much lower than native speaker utterances, which they found to have negative effects on language learners who even adjusted their own utterances in the wrong way. The authors conclude that because of problems in speech melody, synthesized utterances are not useful for foreign language learning. By asking the human speaker to imitate the speech melody to create the stimuli, the authors in Rosenthal-von der Pütten et al.'s study may thus have eliminated the difference between synthesized and natural stimuli that In and Han (2015) identified to be most important. Thus, the fact that Rosenthal-von der Pütten et al. (2016) did not find differences between synthesized and natural speech may be related to the fact that the natural speech did not exhibit natural speech melodies.

Kory-Westlund and Breazeal (2019) investigate the role of the language level a robotic peer uses on four-six year old children's vocabulary learning. They had a robot show a depicted scene on a tablet and tell a short story and then ask the child to tell a story about that scene, after which the robot told a story about a new scene and asked the child again to tell their own story. In the second round, the robot's language level either did, or did not, match the child's language competence level. The two language versions differed in syntax (simple main clauses compared to complex sentences comprising main and subclauses) and more or less complex vocabulary (for instance, basic level versus more specific general language terms). Children came in eight times to play with the robot. Children's vocabulary scores increased more in the condition in which the robot's language matched the child's such that, on average, children in the matched condition picked up almost seven new words, compared to 2.5 in the unmatched condition. These results suggest that the robot's language choices may have an effect on the amount of learning.

To sum up, while the use of robots for language learning is promising and robots have been found to increase children's engagement and interest (at least for a certain amount of time, cf., for instance, Vogt et al., 2019), especially robots as language tutors have not been found to be very effective, and even to be counter-effective in the teaching of native-like speech melodies. However, with robots in other roles than tutors (e.g. in Baxter et al.'s and Kory-Westlund and Breazeal's studies), it seems that speaking style might have an impact. In the current study, we therefore address whether the speaking style of robotic teachers can impact students' performance.

METHODS

In the following, we present a study in which language learners are instructed by a robot who introduces them to the task and the experiment using either a very charismatic or a not so charismatic speaking style. The charismatic speaking style is based on the speech characteristics of Steve Jobs, whereas the other one uses the speech characteristics associated with the speech of Mark Zuckerberg. The two styles have been shown to create different pragmatic effects (e.g., Fischer et al., 2020). The learners are then asked to produce correct interpretations of three questions in English whose intonation contours and stress patterns are noted down in a prosodic notation system (cf. Fischer et al. submitted). Thus, the independent variable in this experiment is the robot's speaking style, and the dependent variables are the errors the participants make when carrying out the task instructed. The focus of the experiment is therefore on the effect of speaking style on students' performance.

Stimuli

The stimuli were created by synthesizing the robot's instructions using the male voice of a free text-to-speech system, and then manipulating them to match the speech characteristics of Steve Jobs and Mark Zuckerberg. The melodic features investigated are those that have previously been found to be related to persuasiveness and positive character traits like enthusiasm, passion, charm, and convincingness in analyses of advertisements and politicians' speeches (Gelinas-Chebat et al., 1996; Rosenberg & Hirschberg 2005, 2009; Biadsy et al., 2007; Nienhuis 2009; Pejčić 2014; Bosker 2021). The acoustic-melodic analysis of various public speakers in Niebuhr et al. (2016) revealed that Steve Jobs' speech features mark one end of the persuasion dimensions whereas Mark Zuckerberg's speech characteristics mark the opposing end of the spectrum among those public speakers investigated. This juxtaposition allowed Niebuhr et al. (2016) to identify potentially influential charismatic speech features, and

TABLE 1 | The acoustic features manipulated.

Acoustic speech feature	Steve Jobs	Mark Zuckerberg
Mean pitch level relative to 100 Hz (st)	8.8	5.4
Mean pitch range (st)	22.9	12.1
Mean acoustic-energy level normalized relative to all instances of "so" (dB)	–3.2	–5
Mean speaking rate (syl/s)	4.4	5.9
Emphatic accent frequency (cpm)	8.4	1.6
Mean silent pause duration (ms)	200	500
Frequency of high-pitched accents (cpm)	17.2	13.8

TABLE 2 | Stress, timing and scaling errors made by students in the Steve Jobs (SJ) and the Mark Zuckerberg (MZ) conditions.

Condition\Error	Stress	Timing	Scaling	Total
SJ	13 (21.7%)	16 (26.7%)	16 (26.7%)	45 (25%)
MZ	15 (25%)	30 (50%)	25 (41.7%)	70 (38.9%)

several studies (Berger et al., 2017) confirm that those speech characteristics are indeed related to charisma, even when used by robots (Fischer et al., 2020).

For the manipulations, we used the PSOLA pitch and duration manipulation functions available in *praat* (Boersma 2001). Changes in acoustic energy were made using Audacity (www.audacityteam.org/). The manipulations resulted in two different versions of the same instructions, one time with the speech characteristics identified for Steve Jobs and another one with those identified for Mark Zuckerberg.

Table 1 provides an overview of the acoustic-melodic parameters manipulated. In general, these features concern the pitch level (measured in semitones relative to a male baseline of 100 Hz, st), i.e., how high or low the fundemental frequency is; the pitch range (measured in semitones, st), i.e., how far up and how far low a given voice moves; the acoustic-energy level (RMS, measured in decibel, dB, and normalized to the dB level of a frequent reference word ("so") in the two speaker's speeches), i.e., how loud the voice is; the speaking rate (measured in syllables per second, syl/s, excluding pauses), i.e., how fast or slow the speech is; the emphatic accent frequency (measured in counts per minute, cpm), i.e., how often a speaker adds expressive accentuation to stressed words; the hesitation frequency (measured in counts per minute, cpm), i.e., how often a speaker uses *uh* and *um*; the duration of silent pauses (measured in deciseconds, ds), i.e., how long the silence lasts; and the frequency of high-pitched accents (measured in counts per minute, cpm). Note that all values in **Table 1** refer to mean values and, thus, to the two speakers' speeches as a whole. Accordingly, we also took them as target values for the robot's utterances as a whole.

When applying these measurements to the robot's utterances, we focused on those acoustic features of the speech signal that were found relevant in the comparison between the different speaking styles of Steve Jobs and Mark Zuckerberg, but we also took into account that the acoustic manipulation would still produce naturally sounding and comparable stimuli. For example, while voice quality is potentially relevant in the perception of the speaker's personality (cf., for instance, Signorello and Demolin 2013), it is difficult to manipulate voice quality, given state-of-the-art resynthesis tools. We therefore restricted the manipulation to the features listed in **Table 1**, which have also been shown to be effective in Fischer et al. (2020), in which the two speaking styles led to different behavioral effects.

During the creation of the robot speech stimuli, the manipulation procedure was conducted iteratively by adjusting each parameter successively. This is necessary because a resynthesis is required after each manipulation before the effect of the manipulation can be evaluated. Thus, the manipulations were applied individually and in as many iterations as necessary to achieve the values described in **Table 1**.

We deemed a manipulation check of our stimuli unnecessary because of extensive previous work that has shown that the two manipulations have significant effects on the speaker's perceived charisma. Specifically, Berger et al. (2017) and Niebuhr (2021b) show in detail that the two speaking styles employed, inspired by Steve Jobs and Mark Zuckerberg respectively, have significantly different effects on the extent to which they are perceived as charismatic. Michalsky and Niebuhr (2019) and Fischer et al. (2020a) furthermore show that the same effect occurs with artificial speakers, like in-car navigation systems and a range of different robots, including Keepons; for instance, in the study involving the Keepon robots, Fischer et al. (2020a) find that the robots that use the speaking style inspired by Steve Jobs to be significantly more passionate, enthusiastic and charming, as well as significantly less boring. We can thus safely assume that the stimuli used in this experiment will yield similar interpretations of the robots as charismatic or not.

The audio files with the instructions were then combined with a video in which a Keepon robot moved slightly as if in coordination with speaking. The text the robot produced was:

> *Hello, we are the Keepons! Thanks for taking the time for this little exercise! We want to teach you how to ask questions in English with the right speech melody. First, my kind human assistant will ask you to fill out a consent form. After that, my human assistant will show you three questions with representations of the speech melody and ask you to record these questions. That's all! Thank you so much already!*

The actual task participants had to fulfill was to produce three questions in English with the appropriate speech melodies. Even though English and Danish are both Germanic languages, the intonation patterns of the two languages are quite different. In particular, in our previous studies (Niebuhr et al., 2017; Fischer et al. submitted), it turned out that native speakers of Danish, who are the participants in our study, have great problems with the production of the rising final tonal gesture in English questions because there is no such rise in questions in Danish; instead, the signaling of speech acts (statements vs. questions) is indicated by the declination of an utterance as a whole (Grønnum 2007: 98). Thus, producing final rises in English questions constitutes a challenge for Danish learners of English as a foreign language. Consequently, we expected our participants to have problems with the task (cf. also Niebuhr et al., 2017).

Participants' task consisted in reading three questions out loud based on a visualization of the speech melody and stress placement of three native-speaker utterances annotated according to ToBI by Hedberg et al. (2017): (10). The examples are thus based on authentic American English questions, more precisely on examples of questions with a low-rise nuclear contour, which Hedberg et al. (2017) find to be the unmarked nuclear contour for *yes-no* questions in their corpus study of American English. We re-interpreted the ToBI notation into drawn intonation contours complemented with stress marking for the prominent syllables (cf. Ladd, 2014). We only made one small adjustment to the third example (Do you still work for a veterinarian) by removing the third, contrastive accent in the sentence because we deemed such a contrastive accent to be confusing without a supporting context. The visualization technique of the intended prosodic realization of these questions had been developed in several experiments and had proven to yield the best results, compared to six other common notation systems (Fischer et al. submitted). **Figure 2** shows the visualizations presented to the students.

The task, to read English sentences based on a visualization of their prosodic realization, is quite difficult for foreign language learners because we do not normally produce intonation contours voluntarily, and because the original English contours feel strange for the native speaker of Danish, where rises in the final intonation contour do not occur. Thus, the task is sufficiently challenging so that ceiling effects are prevented and students are quite likely to fail to some extent. At the same time, it is a realistic and informative task, because students learn to produce English questions in the appropriate way (otherwise risking negative inferences about their personalities, since the transfer of intonation contours from the native to the target language usually leads to unwanted conclusions about a speaker's character cf. Fischer and Niebuhr (2020) concerning the effects of transferring Danish contours into a language in which a final rising contour is expected]. And finally, even though the role of the robot is not to explain how questions in English are to be produced, the task is relatively typical of teaching situations in which teachers provide students with access to resources, like pronunciation dictionaries, that allow them to improve their productions, or ask them to fill out exercising sheets.

Procedure

Forty participants (21 female, 19 male), twenty in each condition, were recruited by three student assistants by approaching them while they were sitting in the common spaces at three campuses of a large Danish university. Participants were all students from a broad range of disciplines and both undergraduates and graduates. Given the prominent role of English in Danish society in general (for instance, panel discussions at prime time may be held in English on Danish TV if international guests are involved), the early introduction of English in school (most often as the first foreign language), and the ubiquity of English at the university in particular, where many courses are taught in English, we can understand all Danish students to be learners of English as a foreign language at an advanced level. The participants were between 18 and 55 years old, with an average age of 26 and a median age of 24.

Participants were given a tablet that played a Powerpoint presentation, where on the first slide they saw a video in which a Keepon robot, the middle robot in a group of three robots (see **Figure 1**), welcomed participants to the experiment and briefly explained the procedure. On the next slide, the students found a link to a consent form, informing them about their right to withdraw from the experiment at any time, and asking them for the permission to record their data, to analyze the data and to publish their data, for instance, at a conference, in separate questions. Then, participants were presented with the visualizations of three English questions (see **Figure 2**), which they were asked to read out loud with the intonation contour and stress pattern visualized. Participants were allowed to practice as often as they like and then provided with an external digital recording device to record their realizations of the three questions.

Data Analysis

In the data analysis, the participants' productions were first annotated using the Kiel Intonation Model (KIM, Kohler 1997; Niebuhr 2021a). The KIM is a phonological intonation model, which analyses utterance intonation in terms of rises, falls and combinations of rises and falls, such as peaks and valleys, on an auditory basis; that is, the analysis is carried out by a prosody expert, in our case, an expert with more than 12 years of prosodic annotation experience. In addition to describing the main intonational movements of the speech melody, the annotation based on KIM also identifies stress

Hello, we are the Keepons! Thanks for taking the time for this little exercise! We want to teach you how to ask questions in English with the right speech melody. First, my kind human assistant will ask you to fill out a consent form. After that, my human assistant will show you three questions with representations of the speech melody and ask you to record these questions. That's all! Thank you so much already!

FIGURE 1 | The Keepon robots used and the text that the robot in the middle presented.

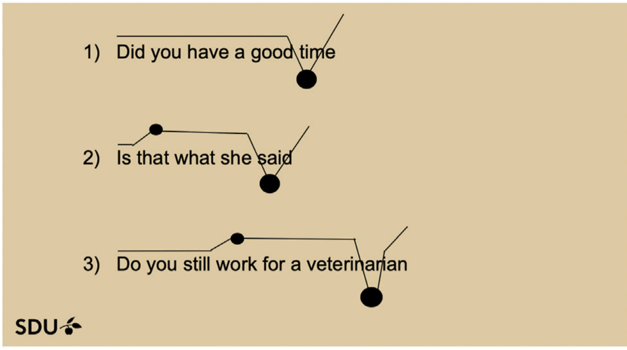


FIGURE 2 | The three questions to be produced.

placement. For instance, regarding our target questions displayed in **Figure 2**, the learners' task is to place the stress on those syllables indicated by the dots in the visualization. The KIM allows the analysis of the placement of the stress, as well as of the right pitch movement. Accordingly, the annotation proceeds in two steps: The first step is to identify pitch-accented words and to distinguish between weak, normal and emphatic prominence levels (cf. Niebuhr et al., 2015; Baumann et al., 2016). The second step is to determine the main melodic movements connected to the stressed words. Based on the annotations of the participants' realizations of the three questions, we analyzed the number and kinds of errors they made. We distinguish between: 1) pitch-timing errors, i.e., errors that occur if the sentence-accent realized by the participant shows a wrong timing (or f0-peak alignment) in relation to the lexically stressed syllable (Niebuhr 2013); 2) pitch-scaling errors, which describe instances in which the pitch movement is different, for instance, if a participant produces a falling contour when a rising contour is indicated (cf. Ladd, 2014); and 3) stress-level errors, i.e., errors that occur if stress is placed on another syllable than indicated. The analysis thus allows us not only to identify the extent to which students' productions are correct, but also what kinds of mistakes they make with regard to the annotation. We are thus interested in how well our participants were able to pronounce the questions as represented by the visualizations. We did not have a panel of native speakers judging the questions in this study because the focus here is not on second-language competence in general, but on the effectiveness of the visualizations in combination with different instructor speaking styles. When we talk about errors, we mean pronunciation errors in relation to the visualizations.

RESULTS

The results show that students who heard the introduction by the robot whose speech was manipulated to match the speech profile of Steve Jobs performed significantly better than those students who heard the introduction from the robot whose speech characteristics matched those of Mark Zuckerberg. Given that each participant produced three questions, there are 60 opportunities for each error type to occur in each condition.

Table 2 summarizes the errors made in the two conditions, and **Figure 3** illustrates their distribution by condition. We carried out a Chi Square test on the data and find a very significant difference between the two conditions on all errors (X^2 (1, $N = 40$) = 7.9858, $p = 0.004715$, $\eta^2 = 0.19$). Furthermore, the differences in errors of timing are very significant (X^2 (1, $N = 40$) = 6.9095, $p = 0.008574$, $\eta^2 = 0.17$), and the difference between the scaling errors approaches significance (X^2 (1, $N = 40$) = 3.0009, $p = 0.083217$, $\eta^2 = 0.08$). Just regarding the stress placement, the difference is non-significant.

Figures 4, 5 illustrate the distributions of the different errors in the two conditions by participant; they show that the different speaking styles affected a large number of participants, and that the effect is not due to a few outliers.

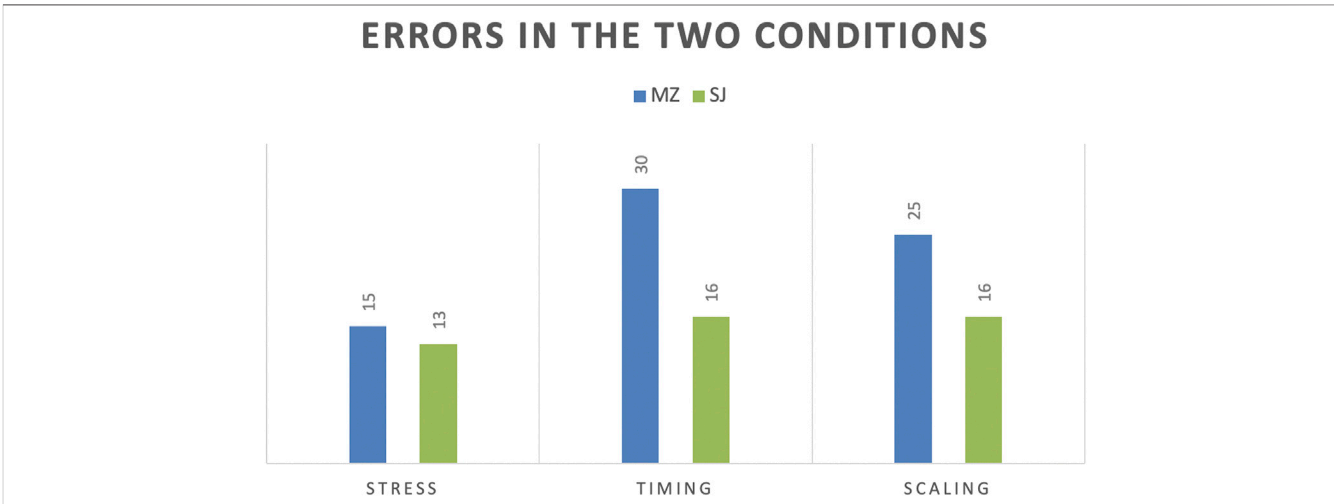


FIGURE 3 | Distribution of the errors made by students in the two conditions.

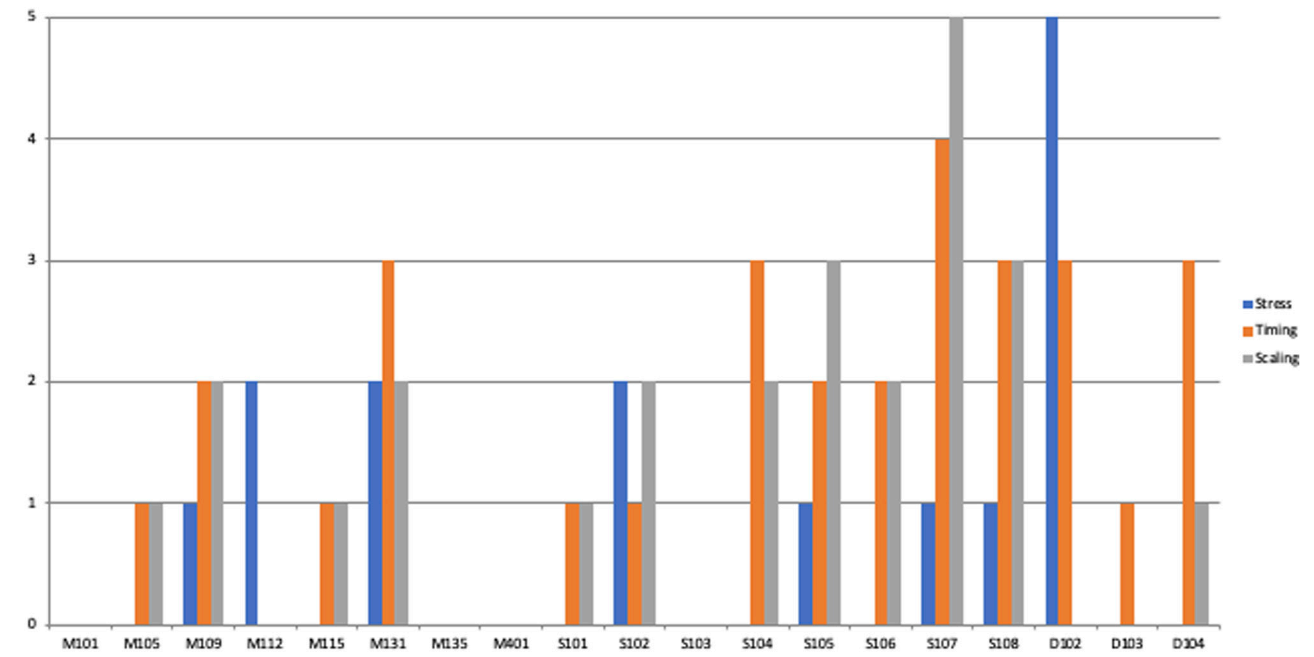


FIGURE 4 | Errors made by the individual participants in the Mark Zuckerberg-condition.

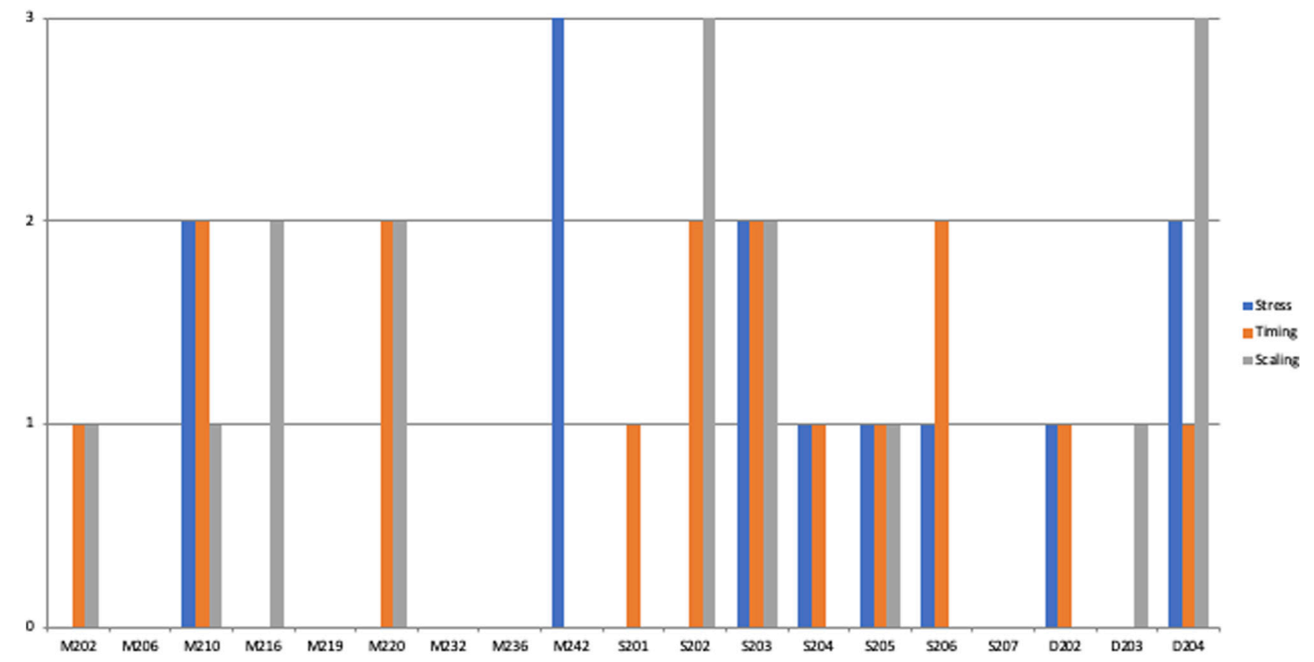


FIGURE 5 | Distribution of errors by participants in the Steve Jobs-condition.

DISCUSSION

The results suggest that the task chosen, to produce English questions with the appropriate prosodic features, was sufficiently difficult, and that even in the condition in which the robot was presenting the instructions with speech characteristics based on Steve Jobs' speech, students still made many errors, including scaling errors that concern the direction of the pitch movement (up or down). Similarly, the timing and placement of stress was challenging. The task was thus adequately difficult but not impossible for the students, and to acquire native-like competence with respect to intonation has been identified as challenging across foreign language teaching research (e.g., Levis 2018). We can conclude that the results were not influenced by potential floor or ceiling effects.

The results obtained furthermore indicate that a teacher's speaking style has a small, but consistent impact on students' objective performance, even if that teacher is a robot. Thus, there is evidence that this part of a teacher's personality significantly influences how well students perform on a given task. While it is still unclear to what extent charisma can be learnt (cf. Abelin 2018), robots can take over the role of providing charismatic instructions in teaching materials.

We suspect that the effect of speaking style on student performance is due to the pragmatic and cognitive effects of charismatic speech in terms of perceived competence, self-confidence and passion. The speech characteristics chosen can be related empirically to these three personality traits, which are likely to cause the effects observed. Specifically, speaking rate and silent pause duration are mostly related to competence (cf. Niebuhr and Michalsky 2019), whereas pitch range and emphatic-accent frequency are related to passion and a higher level of arousal. These in turn can lead to heightened attention and memory in the listener (cf. Niebuhr 2021b). For instance, conveying competence creates trust ("the speaker can do that"), conveying self-confidence creates motivation ("I can do that, too"), and conveying passion creates inspiration and commitment ("I want to do that, too"). This is, we suspect, the reason why a charismatic teacher increases students' performance.

The domain we investigated concerns pronunciation, and thus we have no results on other areas of foreign language learning, such as grammar, vocabulary or interaction. Furthermore, our study focused on adult language learners, and hence the effects may be different for children. Currently, we cannot see a reason why our results should not carry over to other subject areas and other populations, but future work is needed to confirm a general effect.

Another possible limitation may be that the robot only provides the general instructions concerning the different steps involved in the task, so that there is no "teaching" involved in the sense of clarifying contents for a learner. However, we believe that the results are therefore all the more interesting just because the role of the robot as a teacher is so small; if this short introduction to the task already affects the students' performance, then the robot's speech characteristics are likely to affect student performance even more so with a greater involvement of the robot as teacher.

In spite of these potential limitations, we can conclude that robots may serve well as instructors in language teaching, but that their success depends at least to some extent on the speaking style used. Given that most available text-to-speech systems do not provide speaking styles that exhibit characteristics identified as charismatic in previous work, and instead are characterized by features that range low on the charismatic side (see In and Han 2015), this finding draws attention to the need for more adequate speech synthesis for human-robot interaction. Furthermore, our findings suggest that some of the negative findings on robot tutors might actually be due to the respective robot's speaking style - future work will have to shed light on the magnitude of this effect. Finally, the results suggest that speaking style contributes significantly to (robot) teacher personality and student performance.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

KF conceptualized the study, managed and supervised the project and wrote the first draft of the paper. MA created the learning stimuli and contributed the contrastive analysis between English and Danish intonation. ON analyzed the students' speech productions and carried out the statistical analysis. All authors contributed to the content and writing of the final version of the article.

FUNDING

The research presented was carried out with the financial support of the Danish Council of Independent Research under grant number 4180-00359B, which we gratefully acknowledge.

ACKNOWLEDGMENTS

The authors wish to thank Matous Jelínek, Daniela Popa and Selina S. Eisenberger for their help with the data elicitation. We furthermore thank Thomas Buhl and the Centre for Industrial Electronics at SDU for their support with settling the article-processing-charge.

REFERENCES

Abelin, Åsa. C. (2018). Charisma–who Has Charisma and Can I Get it?. *J. Speech Sci.* 6, 2–05.03

Andrist, S., Spannan, E., and Mutlu, B. (2013). "Rhetorical Robots: Making Robots More Effective Speakers Using Linguistic Cues of Expertise," in Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI '13). Piscataway, NJ, USA: IEEE Press, 341–348. doi:10.1109/hri.2013.6483608

Andrist, S., Ziadee, M., Boukaram, H., Mutlu, B., and Sakr, M. (2015). "Effects of Culture on the Credibility of Robot Speech: A Comparison between English and Arabic," in Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15). New York, NY, USA: ACM, 157–164. doi:10.1145/2696454.2696464

Baumann, S., Niebuhr, O., and Schroeter, B. (2016). "Acoustic Cues to Perceived Prominence Levels – Evidence from German Spontaneous Speech," in 8th International Conference of Speech Prosody. Boston, USA, 1–5. doi:10.21437/SpeechProsody.2016-146

Baxter, P., Ashurst, E., Read, R., Kennedy, J., and Belpaeme, T. (2017). Robot Education Peers in a Situated Primary School Study: Personalisation Promotes Child Learning. *PloS one* 12 (5), e0178126. doi:10.1371/journal.pone.0178126

Belpaeme, T., Vogt, P., Van den Berghe, R., Bergmann, K., Göksun, T., De Haas, M., et al. (2018). Guidelines for Designing Social Robots as Second Language Tutors. *Int. J. Soc. Robotics* 10 (3), 325–341. doi:10.1007/s12369-018-0467-6

Berger, S., Niebuhr, O., and Peters, B. (2017). "Winning over an Audience – A Perception-Based Analysis of Prosodic Features of Charismatic Speech," in Proc. 43rd Annu. Meet. German Acoust. Soc. (Daga), Kiel Germany, 1–4.

Biadsy, F., Hirschberg, J., Rosenberg, A., and Dakka, W. (2007). "Comparing American and Palestinian Perceptions of Charisma Using Acoustic-Prosodic and Lexical Analysis," in Proceedings of 8th Interspeech Conference, Antwerp, Belgium, 2221–2224.

Boersma, P. (2001). Praat: A System for Doing Phonetics by Computer. *Glot Int.* 4, 341–345.

Bosker, H. R. (2021). The Contribution of Amplitude Modulations in Speech to Perceived Charisma, in *Voice Attractiveness*. Singapore: Springer, 165–181. doi:10.1007/978-981-15-6627-1_10

De Haas, M., Vogt, P., and Krahmer, E. (2020). The Effects of Feedback on Children's Engagement and Learning Outcomes in Robot-Assisted Second Language Learning. *Front. Robot.* 7, 101ff.

Fischer, Kerstin., Langedijk, Rosalyn., Nissen, L. D., Ruiz Ramirez, Eduardo., and Palinko, Oskar. (2020a). "Gaze-Speech Coordination Influences the Persuasiveness of Human-Robot Dialog in the Wild," in 12th International Conference on Social Robotics, ICSR 2020. United States, Springer, doi:10.1007/978-3-030-62056-1_14

Fischer, Kerstin., Niebuhr, Oliver., Schümchen, Nathalie., and Alm, Maria, Visualizing Intonation for Foreign Language Teaching.

Fischer, Kerstin., and Niebuhr, Oliver. (2020). "Studying Language Attitudes Using Robots," in HRI '20 Companion. March 23–26, 2020. Cambridge, United Kingdom, ACM, doi:10.1145/3371382.3378377

Fischer, K., Niebuhr, O., Jensen, L. C., and Bodenhagen, L. (2020b). Speech Melody Matters-How Robots Profit from Using Charismatic Speech. *J. Hum.-Robot Interact.* 9 (1), 1–21. doi:10.1145/3344274

Gélinas-Chebat, C., Chebat, J.-C., and Vaninsky, A. (1996). Voice and Advertising: Effects of Intonation and Intensity of Voice on Source Credibility, Attitudes toward the Advertised Service and the Intent to Buy. *Percept Mot. Skills* 83, 243–262. doi:10.2466/pms.1996.83.1.243

Grønnum, N. (2007). *Rødgrød Med Fløde: En Lille Bog Om Dansk Fonetik*. Copenhagen: Akademisk Forlag.

Hedberg, N., Sosa, J. M., and Görgülü, E. (2017). The Meaning of Intonation in Yes-No Questions in American English: A Corpus Study. *Corpus Linguistics Linguistic Theor.* 13 (2), 321–368. doi:10.1515/cllt-2014-0020

Jiyoung, I., and Jeonghye, H. (2015). The Prosodic Conditions in Robot's TTS for Children as Beginners in English Learning. *Indian J. Sci. Tech.* 8 (55), 48–51. doi:10.17485/ijst/2015/v8iS5/61476

Klassen, R. M., Durksen, T. L., Al Hashmi, W., Kim, L. E. Ken. Longden., Longden, K., Metsäpelto, R.-L., et al. (2018). National Context and Teacher Characteristics: Exploring the Critical Non-cognitive Attributes of Novice Teachers in Four Countries. *Teach. Teach. Edu.* 72, 64–74. doi:10.1016/j.tate.2018.03.001

Kohler, K. J. (1997). "Modelling Prosody in Spontaneous Speech," in *Computing Prosody. Computational Models for Processing Spontaneous Speech*. Editors Y. Sagisaka, N. Campbell, and N. Higuchi (New York: Springer), 187–210. doi:10.1007/978-1-4612-2258-3_13

Kory-Westlund, J. M., and Breazeal, C. (2019). A Long-Term Study of Young Children's Rapport, Social Emulation, and Language Learning with a Peer-like Robot Playmate in Preschool. *Front. Robot. AI* 6, 81. doi:10.3389/frobt.2019.00081

Ladd, D. R. (2014). *Simultaneous Structure in Phonology*. Oxford: OUP.

Levis, J. M. (2018). *Intelligibility, Oral Communication, and the Teaching of Pronunciation*. Cambridge: Cambridge University Press.

Leyzberg, D., Spaulding, S., Toneva, M., and Scassellati, B. (2012). *The Physical Presence of a RobotTutor Increases Cognitive Learning Gains*. 34th Annual Conference of the Cognitive Science Society, Austin, TX: The Cognitive Science Society.

Lin, S.-H., and Huang, Y.-C. (2016). Examining Charisma in Relation to Students' Interest in Learning. *Active Learn. Higher Edu.* 17 (2), 139–151. doi:10.1177/1469787416637481

Lin, S.-H., and Huang, Yun-Chen. (2017). The Effect of Teacher Charisma on Student Attitude towards Calculus Learning. *Ijsts* 5, 26–32. doi:10.11648/j.ijsts.20170502.12

Niebuhr, O. (2021b). "Advancing Higher-Education Practice by Analyzing and Training Students' Vocal Charisma: Evidence from a Danish Field Study," in 7th International Conference on Higher Education Advances, Spain. 22. Jun 2021. 743-751. doi:10.4995/head21.2021.12827

Niebuhr, O. (2021a). "KIM - the Kiel Intonation Model," in *Prosodic Theory and Practice (Ch. 14)*. Editors J. Barnes and S. Shattuck-Hufnagel (Cambridge: MIT Press).

Niebuhr, Oliver., Alm, Maria., Schümchen, Nathalie., and Fischer, Kerstin. (2017). Comparing Visualization Techniques for Learning Second Language Prosody – First Results. *Int. J. Learner Corpus Res.* 3 (2). doi:10.1075/ijlcr.3.2.07nie

Niebuhr, O., and Michalsky, J. (2019). "Computer-Generated Speaker Charisma and its Effects on Human Actions in a Car-Navigation System Experiment - or How Steve Jobs' Tone of Voice Can Take You Anywhere," in International Conference on Computational Science and its Applications. Cham: Springer, 375–390. doi:10.1007/978-3-030-24296-1_31

Niebuhr, O., Schroeter, B., and Baumann, S. (2015). "Cues to Perceptual Prominence of Prenuclear Pitch Accents – Evidence from German Spontaneous Speech," in Proc. 1st International Conference on Prominence in Language. Cologne, Germany, 1–2.

Niebuhr, O. (2013). "The Acoustic Complexity of Intonation," in *Nordic Prosody XI*. Editors E.-L. Asu and P. Lippus (Frankfurt: Peter Lang), 15–29.

Niebuhr, O., Voße, J., and Brem, A. (2016). What Makes a Charismatic Speaker? A Computer-Based Acoustic-Prosodic Analysis of Steve Jobs Tone of Voice. *Comput. Hum. Behav.* 64, 366–382. doi:10.1016/j.chb.2016.06.059

Nienhuis, M. (2009). "Prosodic Correlates of Rhetorical Appeal: Voice Wave Analysis of Ethos, Pathos and Logos," in Proc. 10th Twente Student Conference. Twente, Netherlands, 1–7.

Nissim, Y., and Simon, E. (2019). The Teacher-Leader's Image: Between Personal Characteristics and Strategies Empowering Learners. *Jel* 8 (3), 136–146. doi:10.5539/jel.v8n3p136

Pejčić, A. (2014). Intonational Characteristics of Persuasiveness in English and Serbian. *Noveaux Cahiers de Linguistique Francaise* 31, 141–151.

Qardaku, N. G. (2019). The Need for Inspirational (Charismatic) Teachers in Today's Classrooms. *Eur. J. Multidisciplinary Stud.* 4 (3), 77–83.

Rosenberg, A., and Hirschberg, J. (2005). "Acoustic/Prosodic and Lexical Correlates of Charismatic Speech," in Proc. 6th Interspeech Conference, Lisboa, Portugal, 513–516.

Rosenberg, A., and Hirschberg, J. (2009). Charisma Perception from Text and Speech. *Speech Commun.* 51, 640–655. doi:10.1016/j.specom.2008.11.001

Rosenthal-von der Pütten, A. M., Straßmann, C., Krämer, N. C., and Krämer, Nicole. C. (2016). "Robots or Agents - Neither Helps You More or Less during Second Language Acquisition," in International Conference on Intelligent Virtual Agents. Cham: Springer, 256–268. doi:10.1007/978-3-319-47665-0_23

Signorello, R., and Demolin, D. (2017). The Influence of Vocal Disorder on the Perception of Charisma in Political Speech. *The J. Acoust. Soc. America* 142 (4), 2641. doi:10.1121/1.5014680

Sutton, Selina. Jeanne., Foulkes, Paul., Kirk, David., and Lawson, Shaun. (2019). "Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, May 2019, 1–14. doi:10.1145/3290605.3300833

Towler, A., Arman, G., Quesnell, T., and Hoffman, L. (2014). How Charismatic Trainers Inspire Others to Learn through Positive Affectivity. *Comput. Hum. Behav.* 32, 221–228. doi:10.1016/j.chb.2013.12.002

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., et al. (2019). "Second Language Tutoring Using Social Robots: a Large-Scale Study," in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 497–505. doi:10.1109/hri.2019.8673077

Conflict of Interest: ON is co-founder of AllGoodSpeakers ApS, a speech-technology company that develops tools and techniques for public-speaker training.

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 in Robotics and AI

ORIGINAL RESEARCH
published: 24 September 2021
doi: 10.3389/frobt.2021.700465

PAL: A Framework for Physically Assisted Learning Through Design and Exploration With a Haptic Robot Buddy

*Soheil Kianzad [1], Guanxiong Chen [2] and Karon E. MacLean [1]**

[1]SPIN Lab, Department of Computer Science, University of British Columbia, Vancouver, BC, Canada, [2]Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

Robots are an opportunity for interactive and engaging learning activities. In this paper we consider the premise that haptic force feedback delivered through a held robot can enrich learning of science-related concepts by building physical intuition as learners design experiments and physically explore them to solve problems they have posed. Further, we conjecture that combining this rich feedback with pen-and-paper interactions, *e.g.*, to sketch experiments they want to try, could lead to fluid interactions and benefit focus. However, a number of technical barriers interfere with testing this approach, and making it accessible to learners and their teachers. In this paper, we propose a framework for Physically Assisted Learning based on stages of experiential learning which can guide designers in developing and evaluating effective technology, and which directs focus on how haptic feedback could assist with *design* and *explore* learning stages. To this end, we demonstrated a possible technical pathway to support the full experience of designing an experiment by drawing a physical system on paper, then interacting with it physically after the system recognizes the sketch, interprets as a model and renders it haptically. Our proposed framework is rooted in theoretical needs and current advances for experiential learning, pen-paper interaction and haptic technology. We further explain how to instantiate the PAL framework using available technologies and discuss a path forward to a larger vision of physically assisted learning.

Keywords: educational robotics, experiential learning, haptic force feedback, interactive drawing, physically assisted learning

OPEN ACCESS

Edited by:
Tony Belpaeme,
Ghent University, Belgium

Reviewed by:
Alexis E. Block,
University of California, Los Angeles, United States
Daniel Davison,
University of Twente, Netherlands

****Correspondence:***
Karon E. MacLean
maclean@cs.ubc.ca

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: *26 April 2021*
Accepted: *31 August 2021*
Published: *24 September 2021*

Citation:
Kianzad S, Chen G and MacLean KE (2021) PAL: A Framework for Physically Assisted Learning Through Design and Exploration With a Haptic Robot Buddy. Front. Robot. AI 8:700465. doi: 10.3389/frobt.2021.700465

1 INTRODUCTION

The learning of topics once delivered in physical formats, like physics and chemistry labs, has moved into digital modalities for reasons from pragmatics (cost, maintenance of setups, accessibility, remote delivery) to pedagogy (topic versatility, personalized learning, expanded parameter space including the physically impossible). Much is thereby gained. However, typically accessed as graphical user interfaces with mouse/keyboard input, these environments have lost physical interactivity: learners must grasp physical concepts in science and math through disembodied abstractions which do little to help develop physical intuition.

Physically interactive robots coupled with an interactive virtual environment (VE) offer an alternative way for students to encounter, explore and collaboratively share and build on knowledge.

While contemporary technology and learning theories have not yet delivered a robot system sufficiently versatile to support a wide range of learning needs and environments, we can nevertheless propose and separately evaluate design dimensions that a haptic robot and accompanying interactive VE enables. The objective of this paper is to facilitate the design and assessment of this new class of learning technology by articulating its requirements *via* a framework.

Experiential learning theorist Kolb (1984) posits a four-phase cycle that learners ideally repeat iteratively: concrete experience (CE), reflective observation (RO), abstract conceptualization (AC), and active experimentation (AE).

In this paper we focus on how a haptic robot might be engaged in the stages of this cycle which naturally lend themselves to physical manipulation: **active experimentation**, through *designing* a virtual experimentation environment suitable for a question they have, and **concrete experience**, through *exploring* the environment they configured.

1.1 A Vision for Physically Assisted Learning: A Sketch-Based Design-Explore Cycle

The ability to draw a model, then feel it (active experimentation around an idea, then associated concrete experience of it—forming and testing a hypothesis) may be key to elevating interactive sketching to experiential learning. When exploring, learners can extend their understanding of a domain of knowledge by physically interacting with a virtual model—making abstract concepts more accessible, and approachable in new ways. When they are designing, physicalized digital constraints combined with sketch-recognition intelligence can help them to expeditiously express their thoughts by sketching to the system, with the added benefit of representing the resulting model to a co-learner. Finally, exploring one's own designs now becomes a holistic cycle: the learner challenges their knowledge by dynamically posing their own questions and mini-experiments as well as others' by designing models, then reflecting on the outcome of interacting with it.

As a concrete example: to "play with" the dynamics of a physical system (*e.g.*, a mass-spring oscillation), a learner is assisted by a force-feedback-enabled drawing stylus to sketch the system on an arbitrary surface. The system recognizes the drawn ink as, say, a mass connected to a ground through a spring. Using the same stylus, the learner can then "grab" the drawn mass and pull on it. To test a question about parallel versus series spring networks, they can mentally predict then quickly draw the two cases and physically compare the result. Similarly, they could test relative oscillatory frequencies by extending the spring then "releasing" it. By writing in a new spring constant ("K = 2") they can modify the spring constant. The same process can be applied in other domains, such as in designing-to-explore an electronic or fluid circuit, and to improvisationally testing equations defining system properties. This use case (**Figure 6**) and others are implemented and elaborated later in this paper.

1.2 Technical Challenges and Ways Around Them

Aspects of the AE and CE experiential learning stages have been studied and validated in isolation using tangible user interfaces, robots and haptic devices, and the results underscore the general promise of this approach (Zacharia et al., 2012; Magana and Balachandran, 2017; Radu et al., 2021). However, few systems support physicalized interaction in both stages, far less fluid transition between them.

This is at least partially due to the technical difficulties of working with present-day versions of these technologies. For example, conventional grounded force-feedback haptic systems can theoretically support VE creation and interaction, but in practice, they require extensive time and expertise not just to create but even to apply small variants in learning purpose, which often is unavailable in a school setting. Their expense, limited-size and desk-tethered workspaces and single-user nature preclude mobility and collaboration and tend to be too high-cost and require significant technical support. Other robot technologies are mobile and collaboration-friendly, but do not convey physical forces—e.g. a robot puck with which a user can control tokens on a graphical screen.

However, a handheld force-feedback tool that combines a spectrum of autonomy with physical interaction can potentially overcome these technical limitations: *e.g.*, a robotic pen which can assist a learner in navigating concepts of physics and math by conveying physical forces modeled by an environment drawn by its holder. Technically, this system must read and understand the user's sketches and notations, translate them into a VE and associated parameterized physical models, then animate this environment mechanically with a physics engine rendered through a suitable force-feedback display—ideally with the same handheld tool with which they drew the environment. A haptic device in the general form of a handheld, self-propelled and high-bandwidth robot can generate untethered, screen-free experiences that encourage collaboration.

This concept is technically feasible today without any intrinsically high-cost elements, with the haptic pen itself fully demonstrated (Kianzad and MacLean, 2018; Kianzad et al., 2020), but significant engineering remains to translate innovations in sketch recognition from other technical domains and integrate them into a full-functioned, low-latency robotic system. Our purpose in this paper is to consider the potential of this approach based on related technology elements as a proxy for a future integrated system which we know is possible to build if proven worthwhile.

1.3 Approach and Contributions

We have designed support based on a theory of activities that has been shown to lead to effective learning, and require this support to meet usability principles suggested by the theory. For example, the cyclical nature of Kolb (1984) et al's learning cycle directs us to minimize cognitive and procedural friction in performing and moving between important cycle activities. Unfettered designing and exploring implies comfortable workspace size and natural command-and-control functions

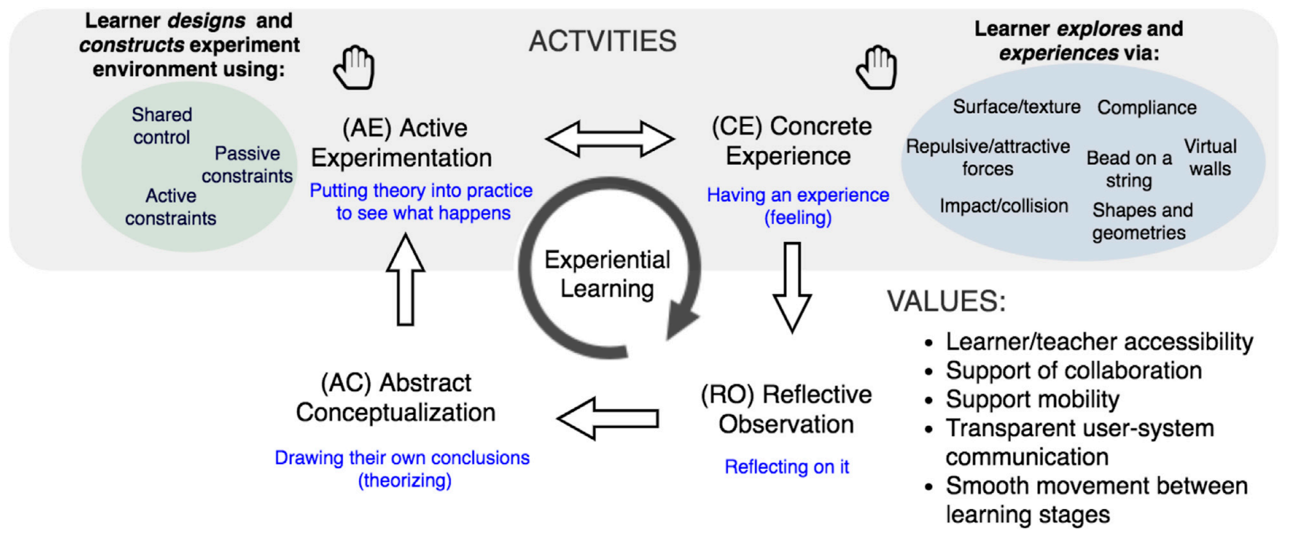


FIGURE 1 | The PAL Framework. Physically Assisted Learning interactions haptically adapt stages of experiential learning from Kolb (1984)'s general framework, with some added features from Honey and Mumford (2000). Hands-on Active Experimentation and Concrete Experience are most amenable to haptic augmentation, enriching the more purely cognitive Reflective Observation and Abstract Conceptualization.

that transfer easily from a student's existing experience—*e.g.*, pen-and-paper diagramming, nomenclature consistent with how they are taught, direct application of parameters, etc. They should not have to switch tools when they switch stages. Meanwhile, their work should be easily visible in a way that teachers and co-learners can see what they are doing and effectively collaborate in their experience (Asselborn et al., 2018; Khodr et al., 2020; Radu et al., 2021).

1.3.1 Getting to Confidence that it Could Work

The scope of this paper is to identify and solve technical obstacles to the instantiation of the theoretically based PAL framework, focusing on the gap in previous work: the connection between physically supported design and explore learning activity, in the form of theoretical rationale and technical proof-of-concept. We need to ensure that the concept's non-trivial realization is feasible, given obstacles ranging from stroke recognition to haptic rendering algorithm and availability of a haptic display with suitable capability and performance.

Only with this evidence will it will be ready to (beyond our present scope) optimize for usability; and thence to evaluate for the pedagogical value of adding physical expression and fluidity to the explore-design-explore cycle. Given the complex and individual process of learning, this will require a sequence of user studies to convincingly validate the framework and its impact on learning gain, as well as generalizablity across multiple platforms.

1.3.2 Guiding Support and Assessing Potentials With an Experiential Learning Framework

We propose a **Physically-Assisted Learning (PAL)** framework through which we can systematically compare different candidate technologies' potentials in *unlocking key activities and values* (**Figure 1**). Through the PAL lens, we view learning *via* the physically supported **activities** of *designing (AE)* and *exploring (CE)*; and assess platforms against key cross-cutting **values** of *learner/teacher accessibility* (Özgür et al., 2017a), support of *collaboration*, untethered (Kianzad and MacLean, 2019), screen-free *mobility*, *transparent* user-system communication (Romat et al., 2019), and *seamless transitioning* between learning stages.

We are using PAL as a tool to understand the impact of device attributes on learning strategies and outcomes, as well as collaborative effectiveness, self-efficacy, creativity, and performance in drawing and design.

Throughout the paper, we will relate needs, technical challenges and approaches to this framework, and consider how the candidate technologies stack up on its values under the two activities of focus.

1.3.3 We Contribute

1) **The Physically Assisted Learning (PAL) framework** which can 1) conceptually and constructively guide the design of haptic science-learning support; and 2) lead directly to articulation of special requirements for *explore*-type contexts like learning, including fluid access to large ranges of model structure and parameterization.
2) **Demonstrations** of 1) means of addressing these needs, for *designing* with innovative application of hand-stroke recognition, and for *exploring* through haptic rendering with a control approach not available in open libraries (namely passivity control); and 2) a technical proof-of-concept system in which *designing* and *exploring* are haptically linked: a user can draw and then feel a virtual environment.
3) **A path forward:** An account of technical considerations, challenges and possible approaches to fully realize this paradigm.

2 BACKGROUND

We introduce past work related to the idea of physicalizing digital manipulatives, relevant classes of haptic force feedback technology, challenges in bringing this kind of technology into education environments, and ways in which haptics have been used for related activities of designing and exploring.

2.1 Adding Physicality to Digital Manipulatives *via* Robots

Physical manipulatives are objects that aid learners in perceiving physics and math concepts by manipulating them, such as the pattern blocks, coloured chips, and coins used in early childhood education to engage learners in hands-on activities. ***Digital manipulatives*** (***DMs***) are physical objects with computational and communication capabilities that can promote different types of thinking in children by engaging them in playing and building. The history of using DMs for education dates to the early 70–80 s in several works from MIT Media Lab's Tangible Media and Epistemology and Learning groups and the Artificial Intelligence Lab. Among them, projects such as Floor Turtle, Graphical Logo, LEGO Mindstorms, Crickets, and Curlybot introduced engaging environments to develop new approaches to thinking about mathematical concepts with encouraging results (Papert, 1980; Resnick et al., 1998; Frei et al., 2000).

Robots are a class of DMs that use motion along with other visual or audio cues to express information. Children can program robots and therefore observe and experience how defining a set of rules results in intentional behaviours in them. This also gives them the freedom to decide what the robot is, based on how the robot behaves. This flexibility potentially helps learners to use the robot as a probe to explore many learning concepts in different contexts (Resnick, 1998).

Haptics can empower digital manipulatives by expanding the imagination beyond the motion of a physical robot, in the behaviour of the virtual avatar and respective feeling of force feedback. While users can manipulate the environment, we posit that the visual and haptic cues can reduce the cognitive load of interpreting the abstract concepts and make the haptic digital manipulative more expressive.

Returning to our mass-spring illustration: a physical mass connected to a real spring is a manipulative that can demonstrate the concepts of elasticity, inertia, vibrations and resonance. A programmable robot can visibly implement the mass-spring behaviour through its reactive motion. With physical user interactivity, this robot becomes a ***haptic digital manipulative***. Combined with a graphical display, it could tangibly render the system with learner-specified parameters—shape, size, spring and mass constants—and expose learners to the reaction forces and dynamics of pulling and bouncing it (Minaker et al., 2016) as well as new combinations of springs, and varying viscosity and gravitational force. Such a system can simulate many other physical systems, *e.g.*, gas, fluid or electronic circuits.

2.2 A Brief Overview of Haptic Force Feedback Technology Relevant to Education

Haptics is the sense of touch, and haptic technology is commonly used to refer to both tactile feedback (*e.g.*, the vibration on your smartphone) and force feedback, which acts on our proprioceptive and kinesthetic senses. Force feedback haptic devices can provide active pushing and pulling; holding the handle of one of these small robots, you can interact with a VE and the dynamics it represents. Force feedback devices come in many forms and capabilities, as portrayed on Haptipedia.org (Seifi et al., 2019).

To support the PAL vision we ultimately want a planar (2D) device that is drawing-friendly, because we see sketching experimental ideas as an intrinsic part of learning. We want large workspace to support big movements, spread out, see one's work; and portability for working in different environments or collaborate around a table without a screen in the way. Cost is crucial to accessibility of any learning technology.

At present, these properties are in substantial conflict. In the interim, to assess technical feasibility we focused on planar (2D) world-grounded force-feedback platforms. This section describes basic terminology, intrinsic tradeoffs, and progress towards this kind of technology we need.

2.2.1 Grounded Force Feedback and Impedance

Force feedback devices require a physical *ground*. Typically, *world*-grounded devices are anchored to a base in order to transfer reaction forces to a ground other than the user's own body, generally *via* links or cables. Device *impedance* is essentially the stiffness it can display. A high-impedance device can strongly resist user movement, either by generating strong actuator forces or by braking and blocking movement. Impedance control is the most common approach to implementing force feedback models: the device is programmed to generate a force in response to a user-imposed positional input. For example, the relation $F = Kx$ describes a simple virtual model of a spring deflection relationship.

2.2.2 Workspace, Mobility and Tethering: Present and Future

With conventional GFF devices, large workspace and impedance range (the difference between minimum and maximum device-renderable impedance) are a tradeoff: 1) minimum renderable impedance *increases* with workspace due to inertia in links and actuators; and 2) maximum renderable impedance *decreases* due to longer and more compliant linkages (Barrow and Harwin, 2008; Zinn et al., 2008). Further, large devices require larger motors, stronger links, better sensors. They are heavy and expensive.

Some low-cost 2D GFF devices targeting education instead go for fabrication ease and low cost. The *Haply* (https://haply.co/) exhibits a "pantograph" configuration, which delivers consistent force over reachable workspace (Gallacher et al., 2016). The Haply has open-source construction guides and a public hAPI software library (Gallacher and Ding, 2017). We chose it for feasibility assessment because of its maturity and support.

In the future, more will be possible. To achieve mobility and large workspace, we need to consider a different approach to grounding than being bolted to a table: hand-held mobile robots which can propel themselves on a surface. Of mobile robots have been used as *active* force feedback displays, two have practical educational potential.

Cellulo is a mobile haptic robot purpose-designed for classroom learning, able to render virtual objects on a 2D-plane (Özgür et al., 2017a). Its omnidirectional, backdrivable mechanism uses a permanent magnet ball to generate vibration-free movement; at 168 g, the size and weight of a large pear, it costs 125 Euros. Cellulo can render variable resistive force feedback and guide users on a specified path or to a certain destination, albeit at a low sample rate and with relatively low magnitude and range of supplied force.

The *Magic Pen* is a low-cost haptic stylus (~50$) which can provide force feedback in 2D (Kianzad and MacLean, 2018). Force grounding is supplied by friction contact between a rolling drive ball and an arbitrary 2D surface, like a tabletop or a vertical whiteboard, able to transmit resistive and guiding forces to the user's hand. Magic Pen supports exploration of a VE, and a later version embeds the drawing capability in an assistive framework for physically assisted manual sketching (Kianzad et al., 2020). The Phasking framework introduces the concept of *control sharing* as well as *bring* and *bound* constraints as a framework to support physically assisted sketching.

Cellulo and Magic Pen are the only plausibly suitable mobile haptic displays of which we are aware which can render force feedback in the large, *i.e.*, an unrestricted workspace. The two differ in many ways: Cellulo is held in a mouse-like grip, and can move autonomously when not held; the Magic Pen is a stylus and thus more suitable for drawing applications, but cannot ambulate on its own. The Magic Pen can deliverable more controllable and larger forces, and is designed to support drawing and sketching whereas Cellulo is physically not suited for this. Purpose-designed for educational purposes, Cellulo as been evaluated in multiple learning scenarios.

2.3 Challenges and Opportunities in Bringing Haptics Into Educational Contexts

Several studies have explored benefits of haptics in education Magana and Balachandran, 2017; Amin et al., 2013; Jones et al., 2006).

We will next discuss the many practical issues in using haptic displays in educational settings, not least the cost of the high-end commercial haptic displays used in these studies.

2.3.1 School Logistics

Besides pedagogical needs, Özgür et al., 2017a propose requirements for a useful educational platform in a class setting: it should be affordable, robust with minimum required initialization, configuration or calibration, and reliable enough to effectively support uninterrupted learning activities. Other considerations drawn from our own classroom work include extreme limits on teacher time (for studies or deployment), technological expertise, and the ability to prioritize such activities. On the technology side, it is hard to justify the expense of sole-use technology, deployment practicalities like batteries and power cords, and the sheer difficulty of students being able to determine when a device is behaving correctly. Classroom sessions are often short, requiring a quick-start system that nevertheless delivers engagement and learning gains right out of the gate (Özgür et al., 2017b).

Validation is a major challenge: it is difficult to validate learning benefit where there are countless variables and controlled studies are not possible. Thus, many studies take a qualitative approach and look for ways in which the haptic modality is changing student strategies, collaboration style, engagement and interest or type of questions (Davis et al., 2017).

2.3.2 Creating Haptically Augmented Learning Environments: Open-Source Haptic Libraries

A multitude of haptic libraries support designers developing haptic VE interactions for a given technology. There are two main categories of functions: rendering haptic behaviors, and connecting the haptic interface modality to other parts of the system and experience, be it an underlying virtual model, graphics and/or sound engines and display, managing other forms of user input and control, and in some cases interaction over a network and with other users and entities. While some have been associated with a specific product, most attempt to generalize support for at least a significant class of devices (*e.g.*, CHAI3d (Conti et al., 2003), hAPI (Gallacher and Ding, 2017)).

Some of these haptic libraries support advanced rendering of complex deformation and collisions both haptically and graphically for sophisticated environments such as surgical training simulations. For educational contexts, we often do not need such complexity. For student-oriented online physics learning materials it is common to see the physical behaviour of an object presented with simplicity *via* an open body diagram and illustration of applied forces (*e.g.*, Perkins et al., 2006).

On gaming platforms, developers use graphic engines to simulate rigid body behavior in a virtual world in procedural animations which move realistically and interactively. Hapticians have exploited game engines for their VE modeling, getting graphic display for free and driving haptic output from the VE simulation; this obviates the need to make or access another physics library for haptic rendering. For example, the hAPI uses a wrapper around the 2D physics simulation library Fisica and turns it into a haptic engine system for educational purposes (Gallacher and Ding, 2017).

However, designing even a simple VE with a library requires basic knowledge of programming and physics, often absent for student or teacher, and often difficult to access in a classroom. Even when a teacher is a technology enthusiast, the uncertainty in predicting learning benefit relative to a large time investment is an understandable barrier. This underscores a broad need for more usable, accessible tools for haptic experience design which go well beyond the need for accessible technology itself.

2.4 Haptics for Designing and Exploring

In this survey of education-related haptics, we focus on the intersection of two primary haptic approaches: 1) haptically rendered virtual environments, and 2) pen-and-paper-based interactions. Although many devices support one or the other, only a few support both features simultaneously.

2.4.1 Design Approaches: Input Methods, Feedback Modalities and CAD Features

Out of many works describing novel input and haptic output, we focus on systems suited to educational applications such as STEM learning and visual art.

VoicePen is a digital stylus that uses non-linguistic vocal, position and pen pressure inputs for creative drawing and object manipulation (Harada et al., 2007). VoicePen uses vowel sounds, variation of pitch, or control of loudness to generate fluid continuous input to the user's pen interactions. *WatchPen* uses a digital stylus, smartwatch and a tablet for drawing inputs while employing vocal and touch input to reduce workflow interruptions, such as tool selection (Hung et al., 2019). These systems' reliance on vocalization make them impractical in classrooms, but they deliver ideas for stylus interactions.

TAKO-Pen is a haptic pen providing pseudo-force-feedback by creating the sensation of sucking on users' fingers through pressure chambers embedded on a handheld surface (Konyo, 2015). *RealPen* is a digital stylus which recreates the sensation of writing on real paper with a pencil through auditory and tactile feedback (Cho et al., 2016). *FlexStylus* allows users to perform tool selection and to draw accurately by bending the pen in various modes (Fellion et al., 2017). Although these novel input and feedback modalities expanded the interaction space between users and haptic devices or digital styluses, they have very specific purposes and do not point to more general sketching tools.

In addition to devices, we looked for innovations in computer-aided drawing (CAD) features for generating engineering or artistic drawings. Parametric sketching is a CAD functionality where users define geometric entities with parameters, and specify relationships between them as constraints: *e.g.*, defining a circle by its central position and radius, or defining two lines as co-linear or of equal length (Pavlidis and Van Wyk, 1985). This function is useful to architects and architects for creating complex architectural or mechanical sketches. Gürel (2019) studied the impacts of parametric drawing with CAD tools on architectural design creation, finding that allowing designers to define parameters and constraints on geometric entities enhanced creative process flexibility. Ullman et al. (1990) emphasized the importance of geometric constraint in CAD tool function for improving clarity in designers' mechanical sketching.

For direct-sketching input, we highlight *ChalkTalk*, which recognizes users' strokes and translates them into meaningful interactions using dynamic visualization and procedural animation to facilitate exploration and communication (Perlin et al., 2018). ChalkTalk is a purely visual medium; we see potential for using its approach when extending PAL-type functionality for more expansive sketch interpretation.

2.4.2 Pen-Based Sketching Tools for Engineering Design and Educational Drawing

In pen-based devices developed for professional and education drawing, we focused on sketching on 2D surfaces to find features that suit PAL needs. In engineering design, *InSitu* provides architects with a stroke-based sketching interface capable of augmenting sites' contextual information from sensor data into sketches and delivering the information *via* pop-ups (Paczkowski et al., 2011). *dePENd* can guide users to draw out shapes (*e.g.*, lines and circles) precisely by providing directional force feedback. It also allows users to deviate from dePENd's guidance so she can edit the shapes at will (Yamaoka and Kakehi, 2013a).

Within educational drawing support for (STEM subjects), most devices were built for sketching math or physics diagrams and equations. *MathPad2* allows users to create animations to represent processes (*e.g.*, a mass block oscillating) in addition to static diagrams or math formula (LaViola and Zeleznik, 2004). *Hands-on Math* places more emphasis on recognizing handwritten math inputs from users and performing calculations such as solving for an unknown variable in an equation (Zeleznik et al., 2010).

While Insitu or MathPad2 support a particular type of drawing, such as architectural sketches, the PAL framework aims to support designers from a wide range of fields—architects, physicists, web developers.

2.5 Relevant Educational Theory and Design Guidelines

2.5.1 Learning Through Experience

In Constructivism, knowledge is seen as deriving from individuals' experiences, rather than as a transferable commodity. Learners actively construct and re-construct knowledge by interacting with the world (Piaget, 1977; Antle and Wise, 2013). According to Piaget's cognitive development theory, to know an object means to act on it. Operation as an essence of knowledge requires the learner to modify and transform an object, and also understand the process of transformation; leading to knowledge of how the object is constructed (Piaget, 1964). Several schools of educators (Montessori and Carter, 1936; Dewey, 1938; Papert, 1980) have emphasized physicality in educational learning tools and direct manipulation of objects. These theories underlie a goal of providing tools that enable learners to operate on multiple instances of knowledge construction.

2.5.2 Extending Experience With Reflection

Meanwhile, Le Cornu (2009) propose three iterative steps of *externalization*, *sense-making of meaning*, and *internalization*, through which reflection links experience to learning. Often discussed in social constructionism literature, these steps have been applied to a wide range of human actions in the world and society, including the use of feedback (from people, or the results of physical "experiments") to develop the meaning of the self.

2.5.3 Haptic Digital Manipulatives as Vehicles for Experience and Reflection

The theories above have been applied to a wide range of tangible user interfaces and digital manipulatives. Through educational robots, experiential learning can be tangible and digitally supported, and specifically invite reflection. Resnick et al., 1998's process of *reflection with robots* starts with the construction of a robot-based environment, in which learners make their own "microworld" by programming it, followed by feedback from robots to help them shape and validate their ideas. Such a reflection cycle can be repeated multiple times, deepening the experience (Frei et al., 2000).

Within early edurobot work, we sought visions for digital manipulatives suitable for more advanced educational topics. We found examples using robots to aid learners in mindful integration or materialization of ideas through the practice of design (Ackermann, 2020); and to support exploration of different domains of knowledge or of abstract concepts by making them more accessible or approachable in new ways (Özgür et al., 2017c).

Instantiating these principles in a digital manipulative could help them to work as an *object-to-think-with*, wherein learners instantiate their ideas into a physical model through the object, and can debug or extend their thinking model regarding the outcome. The process of analyzing the validity of execution motivates learners to think about their own thinking, developing their metacognitive ability. This results in 1) gaining higher-level thinking skills, 2) generating more representations and metaphors for their understanding, 3) improving social communication and a collaborative atmosphere, and 4) forming deeper understanding of the concept among learners (Atmatzidou et al., 2018; Blanchard et al., 2010).

3 A FRAMEWORK FOR PHYSICALLY ASSISTED LEARNING

The motivation for the PAL framework is to exploit benefits postulated above for a haptic digital manipulative, in learning and in pen-and-paper interaction, and turn them into a versatile and effective digital manipulative. We previously introduced Kolb (1984)'s four-stage framework for experiential learning, on which we have based PAL (**Figure 1**). Here, we lay out PAL's theoretical basis, then elaborate on its components and explain how we expect learners and designers to use it.

3.1 Pedagogical Rationale and Components

Learning is iterative: one builds a mental model of a concept by repeatedly interrogating and manipulating a system, forming then testing successive ideas of how it works in a cycle such as Kolb's. Manipulatives are often designed in a way that will support just one part of this cycle—*e.g.*, to create a microworld *or* to directly interact with one.

Our premise is that supporting fluid movement *throughout* the experiential learning cycle will facilitate more resilient mental model formation.

3.1.1 Supporting Kolb's Learning Stages With a Haptic Digital Manipluative

Most of the visions in **Section 2.4**, and the idea of robot-supported reflection more broadly, would support at least one out of Kolb's two "acting in the world" phases: Concrete Experience (CE; having an experience) and Active Experimentation (AE; putting a theory into practice). Here, there is an opportunity for intervention, and also for researchers to observe and try to understand what is happening based on the part of the cycle that is visible. The more internal stages of Reflective Observation (RO; reflecting on an experience) and Abstract Conceptualization (AC; theorizing) are crucial, but can be influenced or inferred only through what happens in the other phases, or through post-hoc assessment, *e.g.*, of changes in conceptual understanding.

The PAL framework's mandate is therefore to help educators focus on physical instruments and strategies that will support learners in CE and AE, and eventually to help us insightfully observe them as they do so.

Early works on edurobots have claimed that robots could be beneficial in all four stages. For example, for Reflective Observation (RO), Resnick et al., 1998 suggested that through its processing power, the robot could speed the reflection cycle—externalizing/internalizing from hypothesis to result; modifying parameters, conditions and even time. For Abstract Conceptualization (AC), Papert (1980) uses gears as an example where learners can use mechanical objects for conceptualizing physics concepts.

Kolb himself argues that the *interaction and manipulation* of tangible objects is an indivisible part of epistemic (knowledge-seeking) exploration, where the learner purposefully changes the learning environment to see its effect and thereby to understand relationships. When suitably framed through availability of multiple perspectives, parameters and factors, manipulation thus might provide at least indirect support for Kolb's Reflecting Observation (RO) stage (Antle and Wise, 2013; Fan et al., 2016).

However, these claims are as yet unsupported. Limited to findings that have been validated in controlled studies, we conjecture that a DM approach's influence on RO and AC will be indirect.

3.1.2 Physically Assisted Learning Components

A useful (that is, versatile) manipulative should be able to provide the basis for productive subsequent reflection and theorizing during both Active Experimentation (AE) and Concrete Experience (CE). Therefore, **we identified *explore* (CE) and *design* (AE) as PAL's key components: activities which a haptic DM must enrich**.

Further support for centering a framework on these two components, as well as clues towards means of implementing them, emerge from other studies of how haptic feedback can support *designing* and *exploring*. Summarizing these, **Table 1** has two features of particular interest. First, we populated it with just two of Kolb's four learning activities, because we found very few examples of attempts to use haptics or other PDMs to directly support reflection or theorizing. Those we did find (*e.g.*, Hallman et al., 2009; Reiner, 2009; Triana et al., 2017) proposed systems or studies whose results either showed no benefit or were inconclusive.

TABLE 1 | Summary of research informing the use and benefits of haptics in learning, organized by the PAL framework's two activity components. [+] indicates a positive benefit, or [–] no added value was found.

Haptic benefits	Design (Active exploration)	Explore (concrete experience)
Understanding and manipulating geometry	[+] Yamaoka and Kakehi (2013b) Drawing accurate geometric shapes [+] Nakagaki and Kakehi (2014) Computer assist collaborative drawing of different shapes [+] Lin et al. (2016) Increasing the passive stylus affordance through haptic guidance	[+] Özgür et al. (2017c) Identifying different shapes and number of edges [+] Minogue and Jones (2009) Understanding the structure and function of the cell membrane transform [+] Jones et al. (2006) Learning morphology and dimensionality of viruses; diagnose mysterious viruses by pushing, cutting and poking
Improving accuracy and speed	[+] Kianzad et al. (2020) Improving accuracy of drawing objects through force feedback assistance [+] Wang et al. (2006) Using haptic feedback in a calligraphy simulation reduces writing errors and improves writing speed [+] Yamaoka and Kakehi (2013b) Drawing accurate geometric shapes	[+] Murayama et al. (2004) Enhancing completion time and interactivity of bimanual tasks [–] Evans (2005) Users were unable to sculpt forms to produce acceptable curved surfaces using haptic feedback [–] Beckers et al. (2020) Haptic human–human interaction does not improve individual visuomotor adaptation
Engagement	[+] Hamza-Lup and Stanescu (2010) Significant increase in students' engagement during the learning activity [+] Younq-Seok et al. (2013) Increasing engagement in word-writing activities [+] Kyung et al. (2007) Increasing confidence and achieving more realistic drawings	[+] Vaquero-Melchor and Bernardos (2019) Enhancing interactions with objects in Augmented Reality [+] Tsetserukou et al. (2010) Providing realistic sensation of physical interaction in a virtual environment [+] Khodr et al. (2020) More engagement in educational robotic activities
Accessibility (e.g., in face of disability)	[+] Mullins et al. (2005) Re-learning to write after a stroke [+] Jafari et al. (2016) Haptics improves task performance of children with physical disabilities (review paper	[+] Wall and Brewster (2003) Allowing visually impaired users to perceive data with greater speed and efficiency
Understanding of underlying concepts	[+] Lopes et al. (2016) Designing an optimum system/model by receiving on-the-go force feedback	[+] Magana and Balachandran (2017) Conceptualizing electrostatic concepts through the sense of touch [+] Zacharia and Michael (2016) Building electrical circuits with one or two bulbs [–] Renken and Nunez (2013) Haptics did not add to learners' ability to understand pendulum principles [+] Zacharia et al. (2012) Understanding mass-beam balance

Secondly, none of the cited studies examined *both* designing and exploring, but treated them as isolated activities. This may have been influenced by the natural affordances of the devices used. For instance, a Haply (in its unmodified state) can be used readily to *Explore*; but to facilitate creation of micro-worlds (*Design*), we felt we needed to hack it—and chose addition of a drawing utensil. In other words, meeting the principles expressed by PAL triggered specific, targeted technology innovation. More is needed to reach the full PAL vision; the framework provides a blueprint to get there.

3.2 Principles for Creating Digital Manipulatives

We assert two overriding principles that guide us in creating versatile digital manipulatives, based on learning theory discussed in **Section 2.5** as well as observations of learners' interactions both with conventional pen and paper and with haptic/robotic devices, across a range of learning scenarios.

3.2.1 A Digital Manipulative Needs to Serve Learners in Expressing Their thoughts (Design)

According to Ackermann (2020), "*To design is to give form or expression, to inner feelings and ideas, thus projecting them outwards and making them tangible*". Design enables individual interactions with and through human made artifacts and involves them in the "world-making" process (Goodman, 1978). The purpose of design goes beyond representing just what exists, by bringing imagination into this existence (Ackermann, 2020).

For example, we often use pen and paper to write down fast-travelling ideas in our minds. Our immediate drawings can reflect our thoughts, experiences and emotions. Particularly for children, drawings reveal the hidden transcripts of their interpretation of the world.

From scribbles to detailed, elaborated productions, sketching is both intellectual play and can help us form, develop and communicate our thoughts, a key part of a conceptual process. Sketching is direct, improvisational, expressive, resists distraction, and may promote deeper cognitive processing. Projecting our ideas onto paper makes our thoughts more tangible, shareable, and justifiable; This enhances our communications with others. A versatile manipulative should work as a medium to exchange information between a user and a computer interactively.

These prior findings and observations support the premise that aid from a suitably configured and supported physical digital manipulative can directly impact the active experimentation phase: specifically, when learners are hypothesizing and planning small tests. The environment altogether should encourage the learner to hypothesize, construct a experimental micro-world and set the conditions for the environment, anticipate the result and test it; and iterate to improve their hypothesis.

3.2.2 A Digital Manipulative Needs to Support Exploration of Domains of Knowledge (Explore)

Two classes of manipulative proposed by Resnick et al. (1998) include *Frobel* Manipulatives (FiMs) to model the world, *i.e.*, provide an intuitive way to experience many concepts in physics by making them more accessible (wooden sphere and cube to feel the natural differences between shapes), and *Montessori* manipulative (MiMs) to model abstract structure—*e.g.*, form an approachable way to make math, and geometry concepts more tangible (golden bead materials used for representing number). Haptics researchers show that even a 1D haptic device can support both of these classes when it works as haptic mirror (Minaker et al., 2016), to mimic physical experience, or as a haptic bridge, connecting a dynamic visualization of a mathematical concept with a haptic representation (Davis et al., 2017). A versatile manipulative should support both classes using physical interaction with the virtual world through force feedback.

Perhaps the most studied aspect of digital and physical manipulative is the role of physicality in simulation learning for concrete experience (CE) stage. Here, learners try out the action and have a new experience. Through physicality, learners can obtain more embodied experiences and perceive information through touch.

3.3 Using the Physically Assisted Learning Framework

3.3.1 Learner's Use

Some examples illustrate PAL's two conceptual activities, wherein a learner constructs a microworld then explores it.

Design: The learner must be able to fluidly express rich information to the system. Assistive force feedback to users' pens while sketching can help them manifest and communicate their ideas to other people and to a computer: it might be more efficient and natural if they can feel virtual constraints that support them in generating smooth curves and straight lines as they draw—on a computer screen, paper, whiteboard or other surface. In the future, we can exploit this design space to empower learners to actively design, make, and change their learning environment based on their hypothesis.

Explore: The tool must provide rich sensory information to the learner. The addition of haptics to a digital manipulative (beyond motion alone) potentially supports a more compelling interpretation so that learners can predict and reason about outcomes based on what they feel as well as see.

In this project we explore these two PAL activities—requisite attributes for an object to think with—along with the connection between them. Although such a device could also be seen as an object to promote computational thinking (Ioannou and Makridou, 2018) we saw it differently. A DM exploits the computational power of the computer to speed up the *learner's* reflection cycle, which leads to more constructive failures (Clifford, 1984). Throughout this process, learners can explore a variety of representations and solution methods. If followed by a consolidation and knowledge assembly stage, together they can create a productive failure process (Kapur and Bielaczyc, 2012).

3.3.2 Education Technology Designer's Use

3.3.2.1 Ideation of Form and Prediction of Haptic Value

Designing technology solutions for learning requires ideating innovative concepts and ideas, but also evaluating and prioritizing them. PAL can help inspire educational technology designers with new ideas, and to understand the potential of adding haptics to a particular domain or context. In addition, our implementation shows a technical example of how to use emerging technological capabilities to solve particular problems.

3.3.2.2 Setting Requirements and Evaluating the Result

PAL can help designers identify *requirements via* experiences that their technology needs to support. Based on **Figure 1**, a designer can create an opportunity map by examining connections between the stages of learning and activity type.

For example, to support collaboration in learning electrostatic forces, a learner can construct the environment (design) by placing the point charges; then invite their partner to experience them (explore). A designer can then focus on finding the haptic controls and feedback which will allow the learner to place the point charges correct places (*e.g.*, equidistant), and how to render the force behaviour as learners move respectively to each other.

Based on these requirements, in *evaluation* a ed-tech designer simply needs (at a first pass) to verify that the requirements are being met when learners interact with the system. Are they able to construct the environment, and then place the charges correct? Can a partner experience this? Is the whole experience engaging and usable enough to invite this kind of collaboration? With the assurance provided by intermediate goal and usability evaluation derived from theory-based guidelines, they will be in a better position to proceed to assess how such a system is influencing learning outcome.

3.4 First Step: Need for a Technical Proof-of-Concept

In past research supporting haptic *design* and *explore* activities (**Table 1**), what is missing is the *connection between* them.

This requires a technical means by which to understand the user's imagination and dialogue in *design* and then bring it into existence by defining its physical, haptic behaviour for *exploring*. For example, if a user draws a microworld consisting of a set of point charges, we need to define the force behaviour of the point charge and make it interactive so that users can feel the forces as they move in the environment.

Once such a system exists, it can misfire for purely technical reasons. For example, expanding the user's available possibilities during *design*—*e.g.*, allowing them to cover a greater variety of concepts in more ways—often introduces new issues such as triggering vibrational instabilities which naturally accompany haptic rendering of dynamic environments with large uncertainties.

In summary, the challenges here are to 1) make an intelligent system that can take unconstrained drawing as an input, and 2) robustly render a wide range of haptic environments with high quality. For the first, advances in artificial intelligence go far in allowing us to infer and display interpretations of user's drawings (Bhunia et al., 2020; Dey et al., 2019). For the second, the field of haptic rendering can contribute advanced control methods which when carefully applied should be able to describe and within bounds, to address the environments that may arise when a user is permitted to create ad hoc environments Haddadi et al., 2015; Diolaiti et al., 2006.

Putting these elements together is, however, a substantial systems-type contribution, and its initial appropriate validation is in technical performance assessment with respect to force and stability outputs relative to known human psychometric capabilities rather than a user study of either usability or learning efficacy. In the following, we will describe and assess performance of our technical-proof-of-concept system which implements this missing, connective aspect of our proposed PAL framework.

4 HAPTICALLY LINKING THE EXPRESSION AND EXPLORATION OF AN IDEA

Currently available processes for generating and modifying content for haptic interaction (**Section 2.3**) impose logistic and cognitive friction between ideation in the form of sketching a problem, idea or experiment the learner would like to understand, and testing that idea in the form of a physicalized model. We aim to reduce this friction.

After describing the technical setup we will use to demonstrate our ideas (overviewed in **Figure 2**), we will work through a series of technical instantiations which support increasingly powerful and wide-ranging cases. Each begins with an education use case illustrating how this level of haptics could be useful. Readers may find the first (rendering a haptic wall) distant from our final goal; we have included it as a means of gradually exposing layers of haptic technology needed to understand more complex implementations. While all of the haptic rendering algorithms described here are well known, we show how they can be combined in new ways with other technical features (*e.g.*, stroke recognition) to meet technical challenges that arise from the requirements of a versatile, unrestricted learning environment.

4.1 Technical Proof-of-Concept Platform: Haply Display and Digital-Pen Stroke Capture

The demonstrations described here use the Haply Robotics' pantograph system (**Figure 3**, https://haply.co/, Gallacher and Ding, 2018) and its hAPI software library (Gallacher and Ding, 2017). The Haply is a low-cost pantograph, relying on 3D-printed parts which together with good-quality motors and fast communication can offer convincing haptic rendering with respect to accuracy, force levels, responsiveness and uniformity across its 14 × 10 cm workspace (https://haptipedia.org/?device=Haply2DOF2016). It communicates sensor and actuator data *via* USB to a VE running on a host computer, typically using the Processing computer language. The hAPI library renders haptic interactions by reading the pantograph's end-effector position (moved by the user's hand) and computing output forces sent to two driver motors.

To capture users' sketch strokes, we used a watermarked paper and a digital pen (Neo Smart Pen, Inc. (2018)) connected to the Haply end-effector. The digital pen captures detailed information about the user's stroke: absolute position, pressure, twist, tilt and yaw. The Neo pen requires watermarked paper, creatable with a standard laserprinter by printing encoded dot files. For erasability and re-usability of sheets, we laminated the watermarked paper and positioned it under the Haply workspace. We calibrated the digital pen's position data with the Haply's encoders. With this system, the user can draw on the laminated paper and the strokes are captured, sent to the host computer and imported to the Processing application that interacts with the Haply.

4.2 Level 1: Rendering Rigid Surfaces and Tunnels

We begin by illustrating how haptics could potentially support learning (in motor coordination) with basic haptic rendering techniques.

4.2.1 Use Case: Handwriting Training Guided by Virtual Walls

Past research on motor training, *e.g.*, post-injury rehabilitation, has elucidated effective strategies for utilizing physical guidance, whether from a human trainer or a programmed haptic appliance. Full guidance of a desired movement does not typically produce good transfer to the unguided case; some studies suggest better results by physically obstructing the desired movement in the face of visual feedback, causing exaggerated motor unit recruitment (Huegel and O'Malley, 2010; Fu et al., 2014). Learning and improving handwriting similarly involves training numerous haptic sensorimotor activities; these employ both fine (fingers) and larger (arms) motor units. It entails significant mental and motor-control practice, particularly for individuals working against special challenges, such as dysgraphia which can impact 25% of the school-aged population (Smits-Engelsman et al., 2001; Guneysu Ozgur et al., 2020).

However, learning and improving handwriting is also a cognitive practice, and often practiced by the young where engagement is also important. Rather than learners comparing their results to a standardized specified outcome, an expert may be able to conceive of better individualized support (more specific, or advanced at a different rate) but requires a means to convey it to the learner as they practice on their own (Gargot et al., 2021).

The priority may thus be easing an expert's customization of exercises, to support repeated self-managed practice (Guneysu

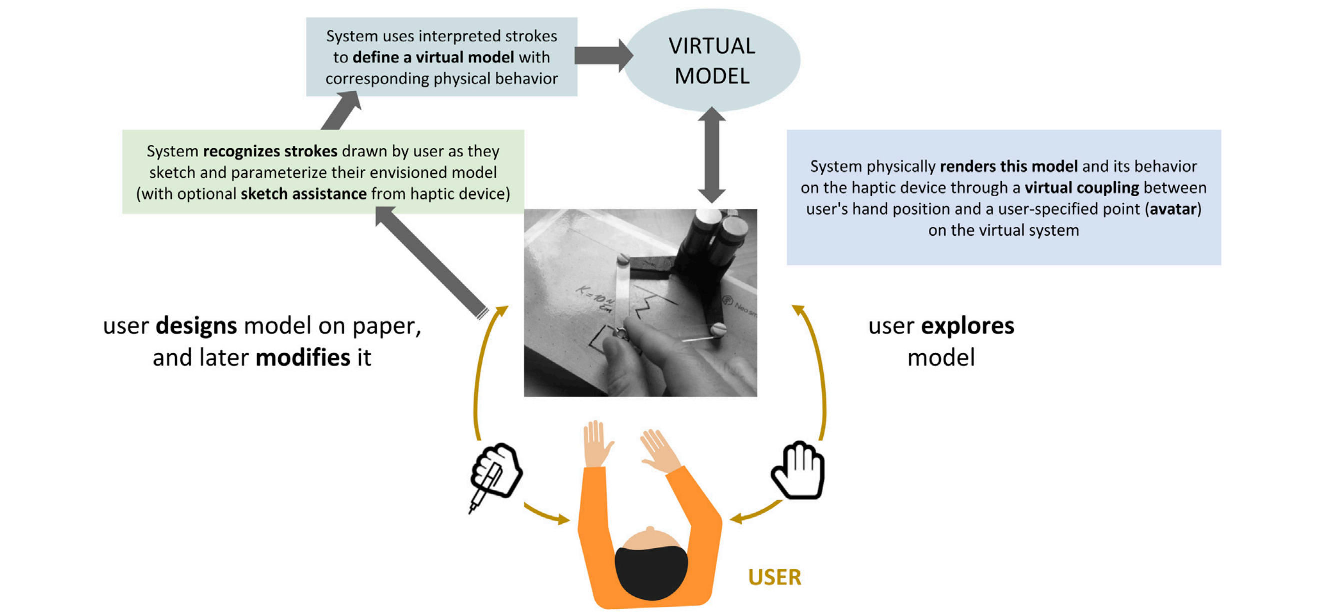


FIGURE 2 | Technical implementation required to support *Design* (green) and *Explore* (blue) learning activities in response to ongoing user input. Details are explained in **Section 4.3.** (The user's graphic from Can Stock Photo, with permission).

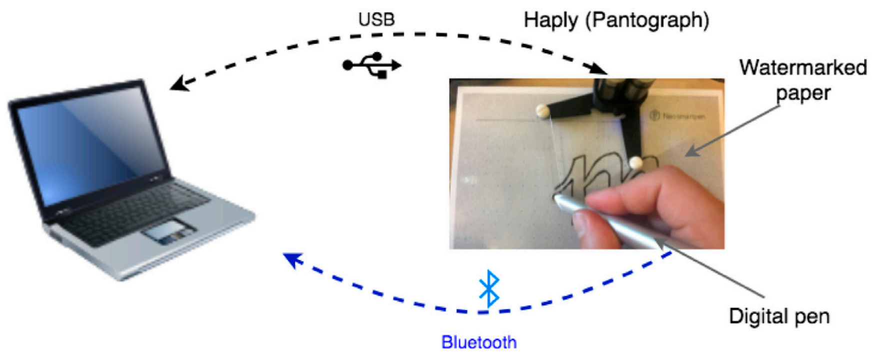


FIGURE 3 | Technical Setup. Our demonstration platform consists of a Haply force-feedback pantograph, a USB-connected digital pen, and a host computer. The Haply communicates position information to the host computer and receives motor commands through a USB port. A digital pen captures and conveys thes user's stroke, along with data opressure, twist, tilt and yaw.

Ozgur et al., 2020). The expert might want to modify details of visual cue presentation and the level and form of haptic guidance (Korres and Eid, 2020; Teranishi et al., 2018); or temporally adapt by reducing force feedback aid over time through control-sharing (Kianzad et al., 2020). Effective feedback must convey correct movements, notify a learner when something goes wrong, and show them how to correct their movement (Asselborn et al., 2018; Amin et al., 2013). Haptic guidance could potentially provide these needed cues when the teacher is not present, without demanding a high cognitive load.

In the PAL framework, the teacher would use the *design* stage, then *explore* to ensure the force feedback works correctly. The learner would access this resource in the *explore* stage.

Here, we show in a basic example targeting elementary school students how a teacher can define a channel within which the learner needs to stay as they trace a letter. This channel will be rendered as a pair of enclosing and guiding haptic walls. This simple demonstration does not attempt best practices for handwriting training, or demonstrate many customization possibilities; it primarily introduces a important building block of haptic rendering, but is also a placeholder for the advanced ways listed above that haptic feedback could be used ti customize handwriting support.

4.2.2 Defining a Wall

There are many ways to define a boundary to a computer program. We require a means that is convenient for a teacher

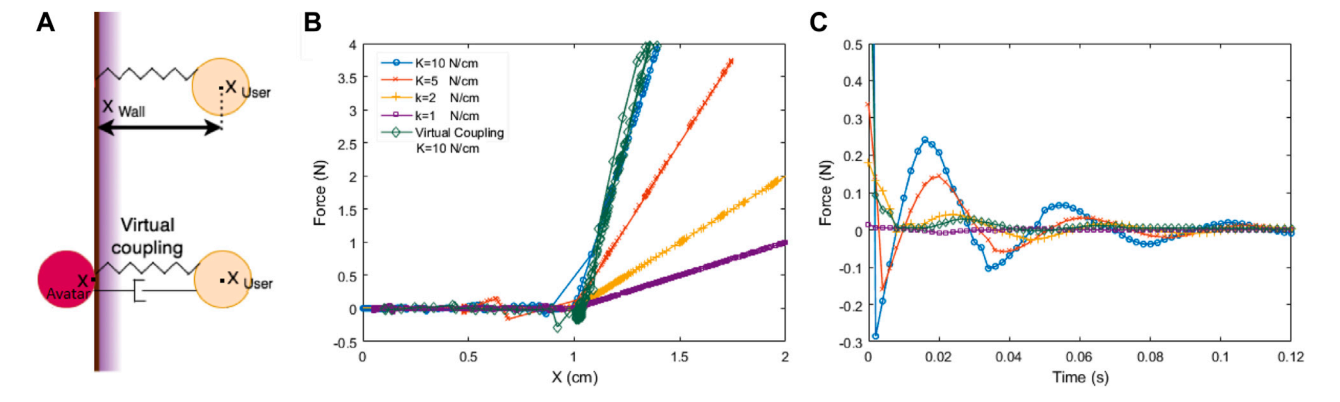


FIGURE 4 | Rendering a haptic wall, using a virtual coupling to achieve both high stiffness and stability. **(A)** Algorithm schematic. **(Upper)** In the simplest rendering method, force depends directly on the distance between the virtual wall and the user's hand (haptic device) as it penetrates the wall: $F = K\,(X_{wall}–X_{user})$. **(Lower)** A *virtual coupling* establishes an *avatar* where X_{user} would be if we could render a wall of infinite stiffness, and imposes a virtual damped-spring connection between X_{user} and X_{avatar}. **(B)** Force-displacement behaviour when the wall is rendered as a direct stiffness or through a virtual coupling. The VC used here also uses the maximum K = 10 *N/cm*, and achieves a similar stiffness as when this K value is used on its own. **(C)** Oscillatory behavior of the conditions from **(B)**. In direct rendering, instability increases with K, but with a VC, a high K is as stable as the softest direct-rendered wall. **(B)** and **(C)** show data sampled from a Haply device.

or therapist. Working in a context of pen-and-paper, we let the teacher sketch the path which they wish the learner to follow. Their strokes are captured as a time-based set of point coordinates. These can be used either directly, if the stroke sample density is adequate, or with a smoothed line fit to them.

We collect the user's strokes as a two-dimensional array, then re-sample it with spatial uniformity and present the result as a one-sided wall. A user can move freely on one side of the wall; if they penetrate the wall from the free direction, they will feel resistance. A teacher can draw a set of one-sided walls as a letter-shaped tunnel to guide a learner in their handwriting practice.

4.2.3 Feeling the Wall: Virtual Coupling

The simplest way to haptically render a wall is to sense the position of the user or haptic device handle, hereafter X_{user}, and compare it with the wall boundary X_{wall}. If X_{user} has penetrated X_{wall}, the penetration distance is multiplied by a stiffness K defining the force that pushes the user out of the wall (**Figure 4A**, upper). However, we typically want to render very stiff walls, while limitations of haptic device force output and sampling rate create a result which is both squishy and unstable (Gillespie and Cutkosky, 1996). As shown in **Figure 4B,C**, increasing K makes a more rigid wall but at the cost of unstable oscillations.

4.2.4 Virtual Coupling for Stiff Yet Stable Walls

An accepted technique for stably rendering stiff walls, *virtual coupling* connects the haptic end-effector position X_{user} to a point representing it in the virtual world which we define as its **avatar** (X_{avatar}, Salisbury and Srinivasan, 1997). A VC links X_{user} to X_{avatar} through a virtual damped-spring, as shown in **Figure 4A**, lower). A stiff VC spring connects the operator more tightly to the virtual model; they can feel more detail, but it can lead to instabilities.

Thus, a VC's parameters (stiffness and damping) need to be tuned to model properties, such as virtual mass and spring magnitudes, device force limits and anticipated interaction velocities. When these are known and constrained to a limited range, a VC can work very well. The VC implementation in the hAPI interface library enables users to change VC parameters (Gallacher and Ding, 2017).

A virtual coupling is closely related to a proportional-derivitive (PD) controller, perhaps the most basic form of automatic control structure. The key goals in tuning either system are to 1) set damping to the minimum needed for stability, to limit energy dissipation and consequently responsiveness; balanced with 2) sufficient stiffness to achieve satisfactorily tight connection to the user's motion. System stability is also challenged when the mass of the virtual entity to which the avatar is either bound or touching is too small, or when the system's update (sampling) rate is slow compared to the dynamics of the system (either the virtual system or the user's movement) (Shannon, 1949).

4.2.5 Wall Performance in Letter-Drawing Use Case

In **Figure 5**, we show the various mechanisms by which a teacher can define and revise a shape which they want a learner to trace (A–D). In (E), we show an example of a learner *exploring* the tunnel defined by the letter outline, including the haptic rendering performance of the virtual coupling as a learner practice to write an *m*. The spring-damper VC filters high frequency force variations and creates smooth guidance as the user slides between and long the walls; the forces keep them within the tunnel. The user's actual position sometimes goes outside the wall, but their avatar remains within it and the learner feels restoring forces pulling them back inside. Depending on velocity, the user position and avatar may be slightly displaced even while within the wall, as the user "pulls" the avatar along through the damped-spring coupling.

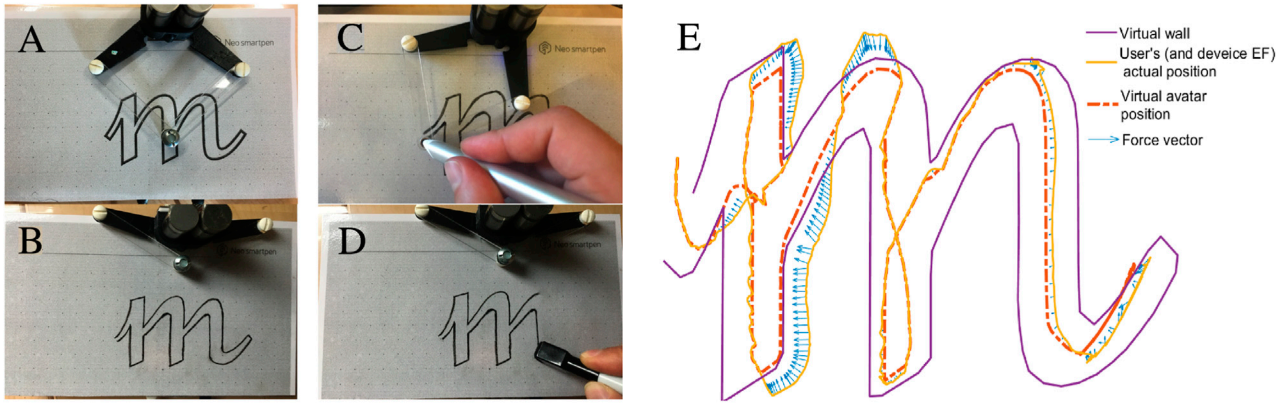


FIGURE 5 | A teacher prepares a handwriting activity by defining a letter shape *m*; the learner will then attempt to form the letter with assisting guidance. To create the *m*, the teacher can **(A)** laser-print a computer-generated graphic on paper, **(B)** draw it by hand, or **(C)** manually draw it with haptic assistance. For erasable media, *e.g.*, pencil on paper or marker on whiteboard, the teacher can **(D)** erase and draw a new exercise. **(E)** Exploring the *m* with ink marks rendered as virtual walls.

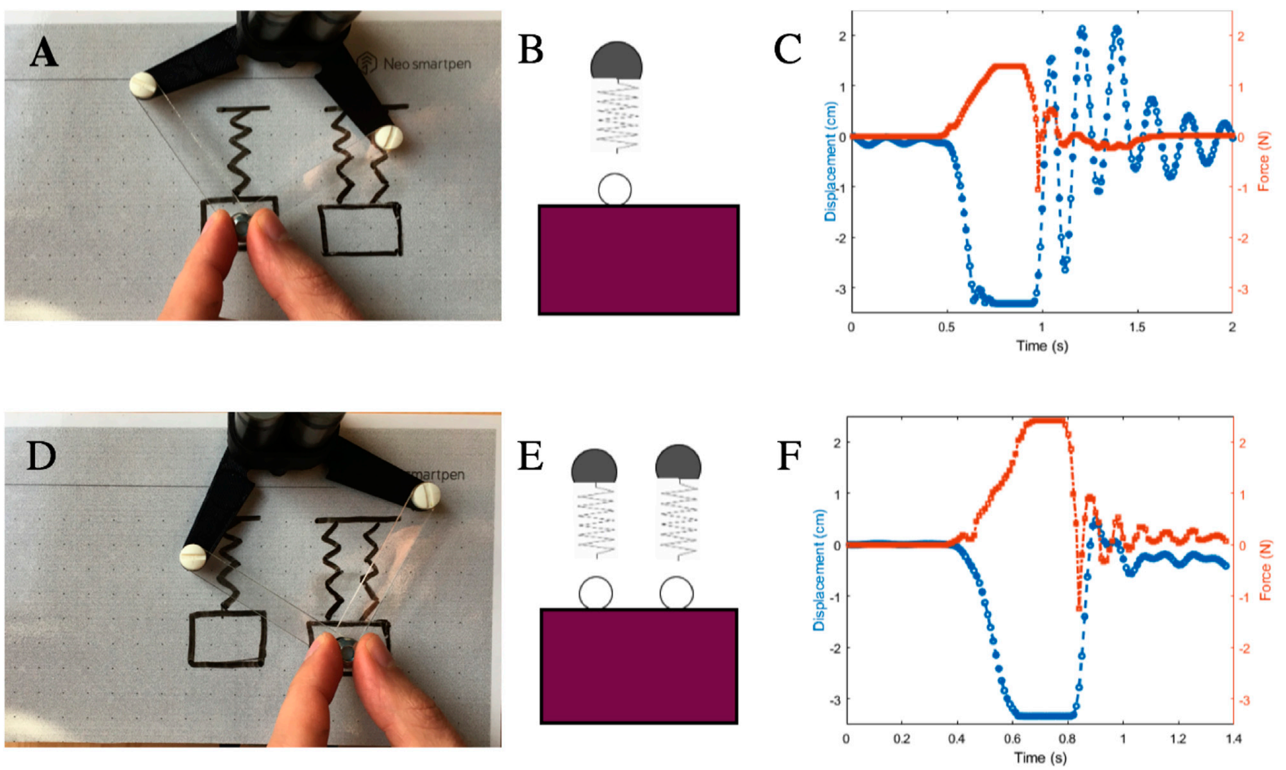


FIGURE 6 | Use case: comparing the dynamic behavior of different spring–mass system configurations by drawing then feeling. **(A)** The user sketches a pair of spring–mass systems using a system-readable notation. **(B,E)** Our system recognizes the user's strokes and incorporates them into virtual models. The user can now "connect" to one of the drawn masses by moving over it and *e.g.*, clicking a user interface button. **(C)** Behavior when connected to the single-spring configuration **(A)**. The system implements the corresponding model **(B)** by pinning X_{avatar} to that mass. The user can then feel the oscillatory force behaviour by "pulling the mass down", extending and releasing the spring. **(D)** The user connects to the two-parallel-springs configuration, and compares its behavior (model E) to the first one. **(F)** compared to **(C)** shows a higher force for the same displacement, and a different oscillatory behavior. This system is implemented using a passivity controller to allow a wide range of M and K values, which are modifiable by hand-writing new values on the sketch.

To extend this example, a teacher could adjust the tunnel width (a step amenable to parameterization) to customize the experience for the learner. The activity can optionally be visualized graphically, or be done entirely on paper. Learner progress can be quantified through statistics of position error (distance between the physical and virtual avatars) and the force magnitude generated in response to this error.

4.3 Level 2: Drawing and Feeling Dynamic Systems

Our second example implements more challenging stroke recognition, and addresses the situation where a virtual coupling is inadequate because of the range of properties that the user may need to access in their design and exploration. The overall flow of the interaction is illustrated in **Figure 2**.

4.3.1 Use Case: a Mass-Spring System

Hook's Law is a linchpin topic in high school physics: along with gravity and friction, students learn about the relation between applied force and the amount of displacement in springs and other stretchable materials. They further must be able to define what a spring constant is, how to compute a net constant assembled through parallel and serial spring assemblies, and with support from their teacher, conduct experiments to verify spring-stiffness hypotheses (Giancoli, 2005). Here, we use a dynamic system consisting of coupled mass and springs to demonstrate the construction of and interaction with a physical system model based on the PAL framework (**Figure 6**).

4.3.2 System Interprets the User's Stroke

We used a 2D recognition library implemented in Processing (the *$1 Unistroke Recognizer* (Wobbrock et al., 2007) to translate user sketches into a virtual model. *$1* is an instance-based nearest-neighbor classifier with a 2-D Euclidean distance function. It can accurately identify 16 simple gesture types, *e.g.*, zigzag, circle, rectangle. To improve performance and customize it to shapes relevant to models our system supports, we created a database to which learners can add their own labeled strokes. In the current implementation, the system starts in a training mode where users draw then type to label their sample; then exit training mode and start designing their experiment.

Our current implementation is modal: it needs to know what kind of a system a user is sketching in order to recognize their marks. A zig-zag could represent a spring in a mechanical system, or a resistor in an electrical circuit. This can be done by manually writing the system type's name on the paper with the digital pen as shown by Lopes et al., 2016—*e.g.*, "Hydraulic lab" triggers a hydraulic simulation. The Tesseract optical character recognizer (OCR) system is one of many robust solutions (Kay, 2007). For simplicity, we selected environments using a graphical user interface.

Reliance on a set notation for sketching has a potential as usability feature or pitfall. If the notation is well known (*e.g.*, taught in the curriculum), it gives the learner a pre-existing language; versus unfamiliar, unmemorable or uncued (*e.g.*, no "tool-tips"). We did not focus on usability refinement at this stage; ensuring it will be an important future step.

4.3.3 System Interprets User Strokes for Model Construction and Parameter Assignment

Ease of environment specification and modification is an important PAL principle. One way that users can specify environment parameters is in the way they draw them. For a mechanical system, a box indicates a mass; mass magnitude is interpreted as the area within the box. Spring stiffness is assigned based on the zigzag's aspect ratio. Haptics can provide assistive guidance to create more accurate drawings. Here, haptic constraints help the user follow implicit geometrical relationships such as relative locations and sizes, through "snapping"; thus the user can perceive when they reach and move beyond the width or length of the previously drawn spring.

Some parameters are harder to indicate graphically, or the user may want to modify an initial value. This could be handled by writing an equation: *e.g.*, set the value of gravitational force with $g = 9.8\ \mathrm{m}/s^2$, or change a spring constant by $K_1 = 10\ N/cm$. As before, recognition can be done with an OCR like Tesseract, a possibility already demonstrated by at least one other system (Lopes et al., 2016).

4.3.4 Unconstrained Experimentation Requires Stepping Up the Control Law

Fluid exploration means that a learner should be able to observe and feel an object's behaviour and reason about it. This requires changing object properties, comparing behaviour between versions of a model and reflecting on the differences.

Above, we introduced the concept of an avatar as key to rendering a wall through a virtual coupling. The avatar's existence was transparent to the user, its existence implicit in their movement. But when we advance to interacting with multiple dynamic systems—to compare them—users must get more explicit with their avatar. To "hold on" to and interact with a part of a virtual model, such as a tool or to probe part of a dynamic system, they must hitch or pin their avatar to that model element, just as they might when selecting a character in a virtual game.

The combined functionality of 1) pinning and unpinning one's avatar to arbitrary system elements, and 2) allowing unconstrained parameter assignment, is a major departure from how a model intended for haptic rendering is typically constructed. Normally, we design an environment with particular components, set its parameters to a pre-known range, and expect the user to interact with it in a particular set of ways—always connecting through a particular avatar linkage. For example, in a surgical simulation, we might have a defined set of tools, and known tissue parameters. Bone and liver have very different properties, and rendering them might be highly complex and computationally expensive, but their properties are known in advance. We can tune a controller (such as a VC) to work with those constrained conditions.

This is no longer the case if parameters can be changed arbitrarily and on the fly, and as usual, the result will be instability. Commonly, several factors can cause instability, such as quantization, delays, and virtual object properties like stiffness and mass. We address this next with the passivity controller.

4.4 Level 3: Expanding the Range of Parameter Exploration Through Passivity Control

To move beyond the simple tuning heuristics above, we reference the notion of *passivity*. A real-world, nonvirtual system like a wood tabletop or mechanical button or doorknob is *energetically passive*—it will never vibrate unstably when we touch, tap or

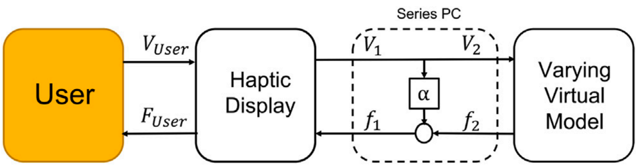


FIGURE 7 | Simulation model of a complete haptic interface system and passivity controller, as implemented here. [Reproduced from Hannaford and Ryu (2002), **Figure 8**—*permission not obtained for review manuscript*]. System blocks are (left to right): user, haptic display, passivity controller α, and virtual environment.

wiggle it because such oscillations require additional energy which they cannot access. The only energy flowing into the interaction comes from our own hand. At best, we can excite a mechanical resonance (*e.g.*, by bouncing a rubber ball, or pumping our legs on a swingset), but this cannot grow in an unlimited way because of the lack of an external energy source.

In contrast, a haptic display is *energetically active:* it accesses an external energy source through its controller. This is necessary for the system to physically simulate a VE's dynamics. However, instability—often manifested as vibrations that grow without bounds, or unnaturally "buzz" upon operator contact—occur when the total energy entering the system from the human operator and the controller's commands is greater than the energy leaving it.

Passivity theory underlies a type of controller which can be designed so as to guarantee stability in systems interacting with humans (Colgate and Brown, 1994; Colgate and Schenkel, 1997; Miller et al., 2004). In essence, passivity controllers bound system movements based on the flow of energy flow through the system: they guarantee overall system passivity by ensuring that the energy input exceeds outputs. It also can achieve global stability through local passivity in subsystems separately. As a result, if we know that other parts of the virtual model and physical device are operating in a passive range, we can focus on the subsystem that the (less predictable) user is interacting with.

4.4.1 Passivity Controller Overview and Design

We designed our *passivity controller* (PC) with the method described by Hannaford and Ryu (2002). In overview (**Figure 7**), the PC is interposed in series between the haptic interface and VE. This location is similar to the virtual coupling controller, and like the VC, the PC works by acting as a virtual dissipative element; the PC differs from a VC through its more targeted energetic accounting system.

The human operator interacts physically with the haptic device in continuous time; however, since the control system is digitally sampled, the VE is computed with a time delay typically specified at 1/10 of the fastest dynamics in the system. The human operator is conceptualized as an *admittance*—a source of *flows* (*i.e.*, movement), and sink of *efforts* (*i.e.*, forces)—and the VE as an *impedance*—a source of efforts and sink of flows.

At the heart of the passivity controller is α, which is in turn based on the *Passivity Observer* (*PO*). The PO, also known as the parameter E_{obsv}, computes the total energy observed in the system at a given moment as:

$$E_{obsv}(n) = \Delta T \sum_{k=0}^{n} f(k)v(k) \tag{1}$$

where ΔT is the sampling time, (f) and (v) are effort and flow (force and velocity) of the 1 port network at time step n. Specifically, f_1 and v_1 are effort and flow for the haptic display, while f_2 and v_2 are for the force computed from the VE computation.

When $E_{obsv}(n)$ is negative, the system in losing energy; for positive values it is generating energy. We compute α as:

$$\alpha(n) = \begin{cases} -E_{obsv}(n)/\Delta T v_2(n)^2, & \text{if } E_{obsv} < 0. \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

After the VE model is updated, its subsystem forces are recalculated, then passed through α before being passed as commands to the haptic display's actuators. f_1, the haptic display command force, is computed as the VE force plus the passivity control component (which is acting to siphon excess energy out of the system).

$$f_1(n) = f_2(n) + \alpha(n)v_2(n) \tag{3}$$

In this implementation

- If the amount of force exceeds the motor force saturation, we subtract the excess amount and add it to the next time step,
- If the user spends significant time in a mode where the PC is active (dissipating considerable energy to maintain stability), energy will accumulate and the PC will not transmit actuation forces until the user has backed away from the dissipation-requiring usage, allowing the PC to discharge. In practice, we reset the PO's energy accumulation to zero every 5 s, scenario-tunable scenario or adapted automatically.

4.4.2 Passivity Controller Performance

4.4.2.1 Example 1: Large-Load Coupling

In our first assessment, we examine the performance of our passivity controller for a simple scenario in which the user's position (X_{avatar}) is "pinned" to a virtual mass as if holding it in their hand. We evaluate performance with two load levels and show how the PCr performs on a large-load coupling.

Virtual Coupling, M = 1X **Figure 8A** shows the displacement (upper) and energy output (lower) of the virtual coupling system of **Section 4.2.3**, *i.e.*, without the PCr. The VC parameters are optimized for this system. Thus, when the user (X_{user}) moves 2 *cm*, X_{avatar} follows smoothly with no overshoot, achieving steady-state by 150 *ms*. The maximum kinetic energy of PC can potentially reduce performance in a normal case where it is not needed, as it may siphon off system energy even when not necessary, being a conservative approach. Therefore for cases close to the system parameters for which the VC was originally tuned, we switch it off.

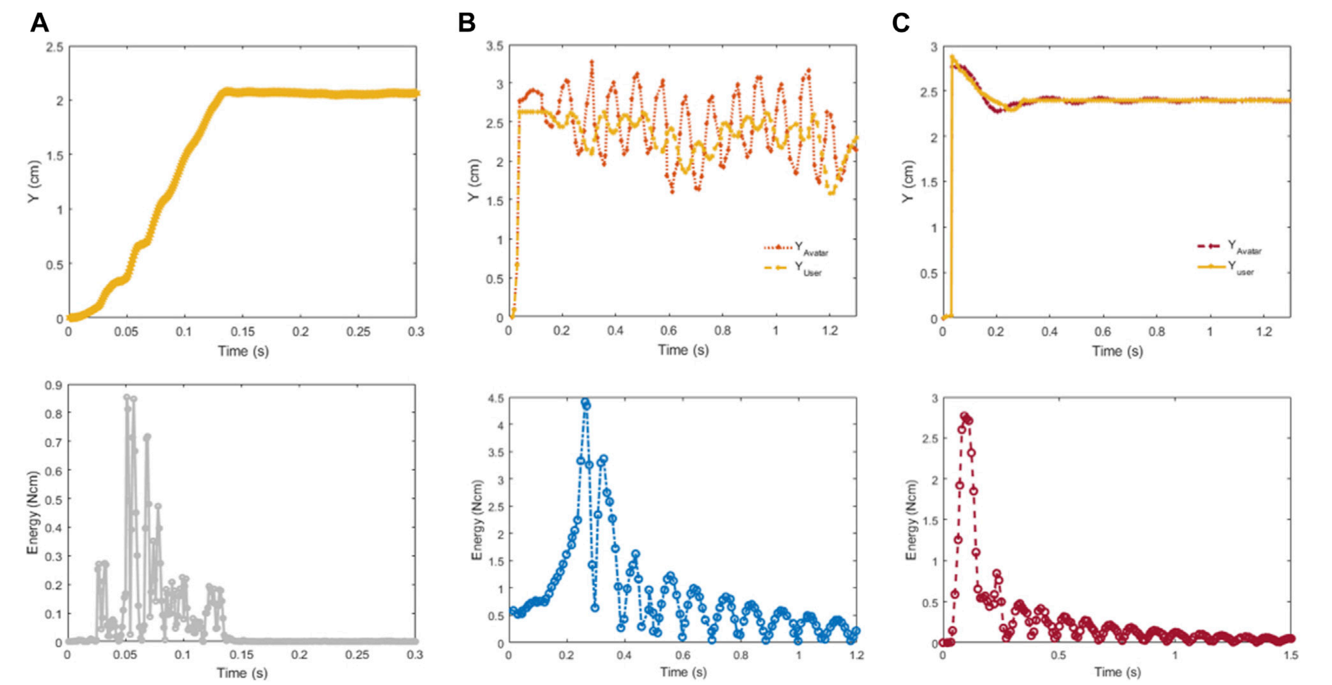


FIGURE 8 | Abrupt movements of varying loads. The position, *i.e.*, X_{avatar} as it tracks X_{user} **(upper)** and kinetic energy **(lower)** of the load for **(A)** the original avatar 25 Gram; **(B)** for the avatar with 20 times more mass than the original avatar without the passivity controller, and **(C)** with the passivity controller.

Virtual Coupling, M = 20X To understand the effect of changing the virtual avatar properties, we investigate a scenario of increasing the mass of the virtual free body being interacted with by 20. **Figure 8B** shows how the system oscillates following the same user movement. Although the oscillation is bounded by physical damping from the user's hand, it can become unstable if the user releases the handle. The system kinetic energy peaks at 4.5 *Ncm* then gradually decreases.

Passivity Control, M = 20X In **Figure 8C**, with the PC active with a large mass, the system overshoots by *44%* but converges within 200 *ms* to the desired displacement. System energy peaks at 2.7 *Ncm* and decreases more quickly than in the VC case for the same mass (B).

4.4.2.2 Example 2: User Interacts With a Virtual Mass-Spring System

The previous example showed how PC can handle a large change in the system's virtual *mass*; how does it do with comparable changes in rendered stiffness as well as the same 20X mass range?

We implement the system as illustrated in **Figure 6**, where a user draws a mass attached to a spring.

Here, **Figure 6B** shows a graphical representation of the recognized mode. Our system recognize a zigzag stroke as a spring and rectangle as a mass, and their connection on the sketch as a kinematic connection between them. The experience is similar to pulling on a real spring: force increases as one pulls further. **Figure 6B** shows the interaction result: as the user pulls down on the spring (change in X_{user}) by around 3 *cm* and then "drops" the force—i.e., stops resisting the haptic display's applied force—the system applies up to 1.27 *N* of force to restore X_{user} to its starting position. The system exhibits a damped oscillation, with two sources: 1) the user's hand and 2) frictions in the haptic display. Here, this is desired behavior faithful to the virtual system dynamics in interaction with the user's hand damping, not a controller instability.

The graphical representation could optionally be displayed to the user to confirm recognition, and animated as they interact. Drawing and animating could implemented on a co-located tablet screen under the haptic display. In future we plan to investigate impacts of employing the user's original strokes versus changing them with a cleaner graphical representation, and of animating the diagrams.

The second row in **Figure 6** shows the user placing two springs in parallel. The learning concept is that springs in parallel sum to a greater combined stiffness than in series, and the operator should feel a tighter (stiffer) connection. In comparision the previous example, the user should perceive a difference in force for the same displacement: the system supplies up to 2.3 *N* force to the user's hand for a similar displacement to the single-spring case. As these results show, this system remains stable under passivity control for a doubling of total stiffness in combination with an already-large virtual mass. This mass-spring example can trivially be extended to include a damper (dissipative element). This is energetically less demanding—a virtual damper does not store energy. In general increasing virtual damping (assuming adequate sampling) reduces susceptibility to large impedance variation (Haddadi et al., 2015).

5 DISCUSSION

We examine this work's contributions, and discuss how the PAL approach can be validated and extended.

5.1 The Physically Assisted Learning Framework, Guidance and Exposed Needs

We drew on general theories of experiential learning to propose a framework that to help haptic and educational experts work together to leverage physical intuition in an effective learning process. This endeavor needs support: learning technology is notoriously hard to evaluate for efficacy, and get feedback on what is helpful. Despite evidence for the role of physical intuition and embodiment in effective learning, we know far less about how to saliently recreate it in digital contexts. Thus, rather than trying to show directly that haptic feedback helps learning, we built on a proven approach in first 1) accepting that *designing* and *exploring* are powerful supports to learning, then 2) seeing how haptic environments can make these activities more powerful than without them.

5.1.1 Metrics

While we have not yet evaluated our technical demonstrations with students, we will in future choose metrics (as per PAL-inspired goals) to highlight how the activities can be more fluid, engaging, focused, intuitive and insightful than without haptics.

5.1.2 Guidelines for Physically Assisted Learning-Inspired Systems

In applying PAL principles we exposed some key requirements. We made progress in translating these to technical challenges, some of which can be addressed with current state-of-art techniques, and others where we need to further innovate. Here we summarize these, noting that while we have identified one pathway to implement them here (**Section 5.2**), we hope that others will find more.

1) *Let learners design their own worlds:* PAL (and experiential learning theory generally) indicates that we should lower friction in letting learners (or in some case their teachers) build their environments. This is an old idea—Scratch and its ilk have born rich fruit—but we need this for environments amenable to haptic display for the purpose of accessing physical tuition.
2) *Let learners explore, iterate and compare those worlds with physical feedback:* Exploration should be informative, flexible and fun. Haptic feedback needs to be clear enough to support insights; it must be possible to jump around easily within an environment and try different things; and the whole process should flow, show insights that might not be otherwise available, surprise and delight. This entails a certain quality of haptic display, and curation of environments (*e.g.*, mechanical systems, electrical, hydraulic, chemistry) that while offering broad scope, also guide the learner on a rewarding path.
3) *Moving between designing and exploring and back should be fluid:* When experiential learning is working as it should, learners will generate more questions as they explore, and want to go back, re-design, compare and ask again. If they have to change modalities or undergo a laborious process to alter the environment or compare different examples, this cycle will be inhibited. We wonder if it is worth trying to stay (graphically) on paper while the digital world plays out through the haptic device, for immersion, focus and the intuitiveness of physical drawing; instead of fussing with a GUI.

5.1.3 Support a Broad Space for Experimentation

Instability is a continual risk for haptic force feedback systems, and could quickly turn anyone off as well as obscuring recognizable physical insights. Tightly restricting the explorable parameter space is an unacceptable solution, since it likewise limits the kinds of experiments to be conducted. Passivity control is one approach to a broader range than the methods currently available to novice hapticians *via* libraries.

5.2 Technical Proof-of-Concept

In the scope of this paper, we have demonstrated at least one full technical pathway for a system that allows a user to design a haptically enabled system by sketching it on paper while adhering to some basic conventions, then interact with that system haptically—and stably—without changing mode or context across a parameter range of which is larger than typically supported in haptic environments. Its and-stroke recognition supports low-friction *designing*, so users can informally sketch ideas, even alter them. For *exploring,* we identified the inadequacy of the conventional rendering method of virtual coupling given the range of system parameters we need to support, and showed how a more specialized controller (based on passivity theory) could take it to this needed level. We encourage curators of haptic libraries to include passivity control support.

5.3 Generalizing to Other Physics Environments: A Bond Graph-Inspired Approach

Our examples demonstrate the ability of a passivity controller to bound a system's energy and prevent instability across a broad range of simulated system parameters. We did this based on a basic mechanical dynamic system, a mass oscillating with different spring combinations. This step can be translated with relative ease to other systems of interest in science learning.

Bond Graph theory (Paynter, 1960; Karnopp et al., 1990) relates physical domains (*e.g.*, mechanics, electronics, hydraulics) based on energetic concepts of *efforts* and *flows*. This commonality is a means to connect domains, but also translate ideas between them. For our purposes, a physical model developed to represent a mechanical system can be translated with relative ease to an electrical domain.

Bond graphs hold threefold value here. First, technically we can exploit its analogies and representation to translate models and their support to other physical domains. Comparable

TABLE 2 | Analogy between some conventional physical domains, reproduced from Borutzky (2011).

Domain	Flow	Effort	Compliance	Resistance	Inertance
Electric	Current	Voltage	Capacitor	Resistor	Inductor
Kinetic translation	Velocity	Force	Spring	Damper	Mass
Kinetic rotational	Angular Velocity	Torque	Torsional Spring	Damper	Inertia
Hydraulic	Flow rate	Pressure	Chamber	Valve	Fluid inertia

properties will be relevant. In Bond graphs, springs (mechanical) and capacitors (electrical) are analogs, both idealized to store energy in the same way, as are mass and inductance, dampers and resistors. **Table 2**, drawn from Borutzky (2011), includes a full list of Bond domain analogies.

Second, these analogs provide a language and convention by which to render physical properties haptically: *e.g.*, effort, flow, resistance and inertance can be developed once and re-used in their relation to one another. It simplifies implementation in new domains.

Thirdly and most interesting pedagogically, these analogs are a powerful way to grasp and generalize fundamental relationships in physical systems. The haptic representation will reinforce this Bond-centered generalization, helping learners to transfer their growing knowledge across domains: once they have mastered how the relations between current, voltage, compliance and resistance work in the electrical domain, they should be able to quickly apply them to kinetic or hydraulic systems. It is often the case that a learner feels more comfortable in one domain; they can use this "home" grounding to support their understanding elsewhere.

5.4 Future Work: The Path Forward

The progress in this paper documents technical feasibility for a basic implementation of a pen-and-paper interaction approach to interactive, self-driven, exploration-centered physical simulation for the sake of learning and gaining physical insight about ideas. Much work remains before we can claim that the concept is ready for roll-out to students and teachers, far less a typical public high school classroom. We lay out some foreseeable next steps.

5.4.1 Validating the PAL Framework: Establishing Impact on Learning

With a theoretically-grounded framework articulated and technical feasibility demonstrated, the next step is to begin confirming the manner and degree to which it actually supports learning. Validating this framework will require a series of focused studies that empirically evaluate the added value of physicality in *design* and *explore* learning phases as well as fluidity in the transition between them. These studies also need to consider factors such as engagement, ownership of knowledge, self-efficacy, self-confidence, and self-paced learning.

One important investigation is the impact of the specific haptic platform: different characteristics (*e.g.*, workspace size, type of grasp, whether they can write as well as feel, ability to propel themselves autonomously, the nature and magnitude of force feedback they can provide) suit them to different usages and learning-environment implementations. Understanding this is the path to generalizing the PAL frameworkt. Meanwhile, PAL itself provides a structure within which we can identify and respond to the advantages and shortcomings of each device as we seek to support specific learning activities.

5.4.2 Basic Access

First and foremost, building haptic worlds and even accessing them interactively requires considerable expertise and infrastructure. Haptic technology is anything but accessible, and this barrier will need to be breached. As for any educational technology, the principle barriers will be in cost, robustness, versatility and usability or expertise.

Low-cost Robust and Highly Portable Technology Lower-cost grounded force feedback devices are becoming more common, but the Haply still costs $300 USD, has a small workspace and is not quite tough enough for a school environment. However, Maker culture has starkly lowered barriers for innovation in this space. With use cases established, we expect to see new GFF device formats be commercialized and toughened. For the ideas described here to succeed, the technology will need to become a commonplace tool. It needs to become a highly portable form factor like a pen or stylus type—such as envisioned in Kianzad et al., 2020.

Versatility There will be myriad ways to use physical interaction in a form factor that one can carry around, perhaps first like the nerdly calculator of the 80 s then becoming more ubiquitously useful as a haptically augmented smartphone stylus.

Usability and Expertise We have called here for lowering friction and barriers to entry for end-users. This also needs to become more true for system designers, allowing them to participate in development from their home discipline and without engineering expertise—*e.g.*, education experts. Input methods, library construction, support groups and other aspects of development ecosystems will move us in this direction.

Eventually, Logistical Deployment with Kids Classrooms are challenging environments. The first point of contact may be science centers and tutoring centers, and potentially on to personal devices (like student calculators) rather than school-supplied technology.

5.4.3 Enhanced Usability, Fluidity and Function

We have described many possible variations and augmentations to a basic implementation, all of which can be explored to discover optimality from logistic and pedagogical standpoints, and inform the direction of further technology development. To name a few (and going beyond innovation in the haptic technology itself):

- CAD-type sketching support at the *design* stage
- More advanced sketch-recognition functions, *e.g.*, setting and modifying simulation parameter values by [re]writing them on paper
- Generating more extensive simulation environments, in multiple domains (*e.g.*, Bond graph extensions)
- Utilizing more sophisticated haptic rendering algorithms as we encounter limits
- Finding good haptic representations for abstract fields such as maths
- Libraries to support educators setting up "design sandboxes"

5.4.4 More Deeply Understanding how Variations Can Maximize Pedagogic Value

Several topics will merit especially deep dives, including longitudinal evaluation; we mention two here.

Paper vs. Digital Boundaries When should the system stay entirely paper-based, and when move between graphical and paper—what are the value and limitations of each modality in the situation where for the first time we have a choice about it?

Supporting Collaboration How do we best exploit the power of collaborative learning, allowing students to share ideas with peers and jointly work out problems? This sketch-based approach paired with personal or shared haptic devices could be extended to remote learning scenarios with the use of touchscreen tablets. We anticipate that haptic feedback can provide a new vocabulary through which students can communicate.

6 CONCLUSION

A long-awaited promise of ubiquitous computing (Weiser, 1991) is natural access to computational power where and when we need it. Yet, for the most part we remain tied to a small screen and a keyboard or tablet, with constrained space to work, keystroke input, a single viewport with many distractions, and interaction generally on the terms of the device.

In this paper we proposed an approach to support multimodal learning with potential benefits to embodied learning and thinking. It includes a framework drawn from validated theories of experiential learning translated to the physical domain to guide system designers in creating educational systems focused on *designing* and *exploring*; underscoring of the importance of fluid, same-modality movement between these learning phases; demonstrations of the technical feasibility of implementing both idea capture and physical rendering in a pen-and-paper environment; and guidelines and assessment of how to move such a vision forward. We demonstrated these ideas on a fixed small-workspace device, but untethered, infinite workspace grounded force feedback has been prototyped and could be commercially viable given demand.

The present work points to a path away from tethered, disembodied interaction, examining ways to harness the natural fluidity and ease of pen-and-paper interactions and connect them to powerful digital simulation for the purpose of simulation, gaining physical, embodied insight, problem solving and thinking with our sense of touch as well as our heads and eyes. A graphical viewport is not always needed when we have our imagination, a sketchpad and hands to feel.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

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

REFERENCES

Ackermann, E. K. (2020). 26 Experiences of Artifacts. *Designing Constructionist Futures: Art, Theor. Pract. Learn. Designs.*, 275. doi:10.7551/mitpress/12091.001.0001

Amin, M. M., Zaman, H. B., and Ahmad, A. (2013). "Visual Haptic Approach Complements Learning Process of Jawi Handwriting Skills," in 2013 5th International Conference on Information and Communication Technology for the Muslim World, Rabat, Morocco (New York, NY: IEEE), 1–6. doi:10.1109/ICT4M.2013.6518888

Antle, A. N., and Wise, A. F. (2013). Getting Down to Details: Using Theories of Cognition and Learning to Inform Tangible User Interface Design. *Interacting Comput.* 25, 1–20. doi:10.1093/iwc/iws007

Asselborn, T., Guneysu, A., Mrini, K., Yadollahi, E., Ozgur, A., Johal, W., and Dillenbourg, P. (2018). "Bringing Letters to Life," in Proceedings of the 17th ACM Conference on Interaction Design and Children (New York, NY, USA: ACM), 219–230. doi:10.1145/3202185.3202747

Atmatzidou, S., Demetriadis, S., and Nika, P. (2018). How Does the Degree of Guidance Support Students' Metacognitive and Problem Solving Skills in Educational Robotics?. *J. Sci. Educ. Technol.* 27, 70–85. doi:10.1007/s10956-017-9709-x

Barrow, A. L., and Harwin, W. S. (2008). "High Bandwidth, Large Workspace Haptic Interaction: Flying Phantoms," in 2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 295–302. doi:10.1109/haptics.2008.4479961

Beckers, N., van Asseldonk, E. H. F., and van der Kooij, H. (2020). Haptic Human-Human Interaction Does Not Improve Individual Visuomotor Adaptation. *Sci. Rep.* 10, 19902–19911. doi:10.1038/s41598-020-76706-x

Bhunia, A. K., Das, A., Muhammad, U. R., Yang, Y., Hospedales, T. M., Xiang, T., et al. (2020). Pixelor. *ACM Trans. Graph.* 39, 1–15. doi:10.1145/3414685.3417840

Blanchard, S., Freiman, V., and Lirrete-Pitre, N. (2010). Strategies Used by Elementary Schoolchildren Solving Robotics-Based Complex Tasks: Innovative Potential of Technology. *Proced. - Soc. Behav. Sci.* 2, 2851–2857. doi:10.1016/j.sbspro.2010.03.427

Borutzky, W. (2011). *Bond Graph Modelling of Engineering Systems*. New York, NY: Springer.

Cho, Y., Bianchi, A., Marquardt, N., and Bianchi-Berthouze, N. (2016). "Realpen," in Proceedings of the 29th Annual Symposium on User Interface Software and Technology. doi:10.1145/2984511.2984550

Clifford, M. M. (1984). Thoughts on a Theory of Constructive Failure. *Educ. Psychol.* 19, 108–120. doi:10.1080/00461528409529286

Colgate, J. E., and Brown, J. M. (1994). "Factors Affecting the Z-Width of a Haptic Display," in *IEEE International Conference on Robotics & Automation* (San Diego, CA), 3205–3210.

Colgate, J. E., and Schenkel, G. G. (1997). Passivity of a Class of Sampled-Data Systems: Application to Haptic Interfaces. *J. Robotic Syst.* 14, 37–47. 10.1109/ACC.1994.735172.

Conti, F., Barbagli, F., Balaniuk, R., Halg, M., Lu, C., Morris, D., et al. (2003). The CHAI Libraries. *Proc. Eurohaptics.* 2003, 496–500.

Davis, R., Martina-Ortiz, M., Schneider, O., MacLean, K. E., Okamura, A. M., and Blikstein, P. (2017). "The Haptic Bridge: Towards a Theory of Haptic-Supported Learning," in 16th International Conference on Interaction Design and Children, 51–60.

Dewey, J. (1938). *Logic: The Theory of Inquiry*. New York, NY: Holt and company

Dey, S., Riba, P., Dutta, A., Llados, J., and Song, Y.-Z. (2019). "Doodle to Search: Practical Zero-Shot Sketch-Based Image Retrieval," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2179–2188. doi:10.1109/cvpr.2019.00228

Diolaiti, N., Niemeyer, G., Barbagli, F., and Salisbury, J. K. (2006). Stability of Haptic Rendering: Discretization, Quantization, Time Delay, and Coulomb Effects. *IEEE Trans. Robot.* 22, 256–268. doi:10.1109/TRO.2005.862487

Evans, M. A. (2005). Rapid Prototyping and Industrial Design Practice: Can Haptic Feedback Modelling Provide the Missing Tactile Link?. *Rapid Prototyping J.* 11, 153-159. doi:10.1108/13552540510601273

Fan, M., Antle, A. N., and Cramer, E. S. (2016). "Exploring the Design Space of Tangible Systems Supported for Early reading Acquisition in Children with Dyslexia," in Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction - TEI '16 (New York, New York, USA: ACM Press), 689–692. doi:10.1145/2839462.2854104

Fellion, N., Pietrzak, T., and Girouard, A. (2017). "FlexStylus," in Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (New York, NY, USA: Association for Computing Machinery), 375–385. doi:10.1145/3126594.3126597

Frei, P., Su, V., Mikhak, B., and Ishii, H. (2000). "Curlybot," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA: ACM), 129–136. doi:10.1145/332040.332416

Fu, M. C., DeLuke, L., Buerba, R. A., Fan, R. E., Zheng, Y. J., Leslie, M. P., et al. (2014). Haptic Biofeedback for Improving Compliance With Lower-Extremity Partial Weight Bearing. *Orthopedics.* 37, e993–8. doi:10.3928/01477447-20141023-56

Gürel, A. (2019). Cognitive Comparison of Using Hand Sketching and Parametric Tools in the Conceptual Design Phase. Ankara: Bilkent University

Gallacher, C., and Ding, S. (2017). hAPI Library. Available at: https://github.com/haplyhaptics

Gallacher, C., and Ding, S. (2018). "The Haply Development Platform: A Modular, Open-Source Haptic Ecosystem that Enables Accessible Multi-Platform Simulation Development," in *IEEE Haptics Symposium (HAPTICS)* (San Francisco: IEEE), 156.

Gallacher, C., Mohtat, A., Ding, S., and Kövecses, J. (2016). "Toward Open-Source Portable Haptic Displays With Visual-Force-Tactile Feedback Colocation," in 2016 IEEE Haptics Symposium (HAPTICS) (IEEE), 65–71. doi:10.1109/haptics.2016.7463157

Gargot, T., Asselborn, T., Zammouri, I., Brunelle, J., Johal, W., Dillenbourg, P., et al. (2021). "it Is Not the Robot Who Learns, it Is me." Treating Severe Dysgraphia Using Child–Robot Interaction. *Front. Psychiatry.* 12, 5. doi:10.3389/fpsyt.2021.596055

Giancoli, D. C. (2005). *Physics: Principles with Applications*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Gillespie, B., and Cutkosky, M. (1996). Stable User-Specific Haptic Rendering of the Virtual wall. In ASME International Mechanical Engineering Conference and Exposition, Symposium on Haptic Interfaces (Atlanta, GA: ASME Press). 58, 397–406.

Goodman, N. (1978). *Ways of Worldmaking*. Indianapolis, IN: Hackett Publishing Company.

Guneysu Ozgur, A., Özgür, A., Asselborn, T., Johal, W., Yadollahi, E., Bruno, B., et al. (2020). Iterative Design and Evaluation of a Tangible Robot-Assisted Handwriting Activity for Special Education. *Front. Robot. AI.* 7, 29. doi:10.3389/frobt.2020.00029

Haddadi, A., Razi, K., and Hashtrudi-Zaad, K. (2015). Operator Dynamics Consideration for Less Conservative Coupled Stability Condition in Bilateral Teleoperation. *Ieee/asme Trans. Mechatron.* 20, 2463–2475. doi:10.1109/tmech.2014.2385637

Hallman, G., Paley, I., Han, I., and Black, J. B. (2009). "Possibilities of Haptic Feedback Simulation for Physics Learning," in EdMedia+ Innovate Learning (Waynesville, NC: Association for the Advancement of Computing in Education (AACE)), 3597–3602.

Hamza-Lup, F. G., and Stanescu, I. A. (2010). The Haptic Paradigm in Education: Challenges and Case Studies. *Internet Higher Education.* 13, 78–81. doi:10.1016/j.iheduc.2009.12.004

Hannaford, B., and Jee-Hwan Ryu, J.-H. (2002). Time-Domain Passivity Control of Haptic Interfaces. *IEEE Trans. Robot. Automat.* 18, 1–10. doi:10.1109/70.988969

Harada, S., Saponas, S., and Landay, J. A. (2007). "Voicepen: Augmenting Pen Input With Simultaneous Non-Linguistic Vocalization," in Proceedings of the 9th International ACM Conference on Multimodal Interfaces (ACM).

Honey, P., and Mumford, A. (2000). *The Learning Styles Helper's Guide*. Maidenhead: Peter Honey Publications.

Huegel, J. C., and O'Malley, M. K. (2010). "Progressive Haptic and Visual Guidance for Training in a Virtual Dynamic Task," in *IEEE Haptic Symp HAPTICS '10* (Waltham, MA: IEEE Press). doi:10.1109/haptic.2010.5444632*)*

Hung, M., Ledo, D., and Oehlberg, L. (2019). "WatchPen," in Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (New York, NY, USA: Association for Computing Machinery). doi:10.1145/3338286.3340122

Inc, N. S. (2018). The Neo Smart Pen. Available at: https://www.neosmartpen (Accessed March 9, 2016).

Ioannou, A., and Makridou, E. (2018). Exploring the Potentials of Educational Robotics in the Development of Computational Thinking: A Summary of Current Research and Practical Proposal for Future Work. *Educ. Inf. Technol.* 23, 2531–2544. doi:10.1007/s10639-018-9729-z

Jafari, N., Adams, K. D., and Tavakoli, M. (2016). Haptics to Improve Task Performance in People With Disabilities: A Review of Previous Studies and a Guide to Future Research With Children with Disabilities. *J. Rehabil. Assist. Technol. Eng.* 3, 2055668316668147. doi:10.1177/2055668316668147

Jones, M. G., Minogue, J., Tretter, T. R., Negishi, A., and Taylor, R. (2006). Haptic Augmentation of Science Instruction: Does Touch Matter?. *Sci. Ed.* 90, 111–123. doi:10.1002/sce.20086

Kapur, M., and Bielaczyc, K. (2012). Designing for Productive Failure. *J. Learn. Sci.* 21, 45–83. doi:10.1080/10508406.2011.591717

Karnopp, D. C., Margolis, D. L., and Rosenberg, R. C. (1990). *System Dynamics: A Unified Approach*. New York: Wiley.

Kay, A. (2007). Tesseract: An Open-Source Optical Character Recognition Engine. *Linux J.*, 2.

Khodr, H., Kianzad, S., Johal, W., Kothiyal, A., Bruno, B., and Dillenbourg, P. (2020). "Allohaptic: Robot-Mediated Haptic Collaboration for Learning Linear Functions," in 2020 29th IEEE International Conference on Robot and Human Interactive Communication, Naples, Italy (New York, NY: IEEE), 27–34. doi:10.1109/ro-man47096.2020.9223563

Kianzad, S., Huang, Y., Xiao, R., and MacLean, K. E. (2020). "Phasking on Paper: Accessing a Continuum of Physically Assisted Sketching," in ACM Conference on Human Factors in Computing Systems (CHI'20) (Honolulu: ACM Press), 1–11. doi:10.1145/3313831.3376134

Kianzad, S., and MacLean, K. E. (2019). "Collaborating through Magic Pens: Grounded Forces in Large, Overlappable Workspaces," in *Haptic Interaction*. Editors H. Kajimoto, D. Lee, S.-Y. Kim, M. Konyo, and K.-U. Kyung (Singapore: Springer Singapore), 233–237. doi:10.1007/978-981-13-3194-7_54

Kianzad, S., and MacLean, K. E. (2018). "Harold's Purple Crayon Rendered in Haptics: Large-Stroke, Handheld Ballpoint Force Feedback," in 2018 IEEE Haptics Symposium (HAPTICS), 106–111. doi:10.1109/HAPTICS.2018.8357161

Kolb, D. (1984). *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, NJ: Prentice-Hall.

Konyo, M. (2015). "Tako-pen: A Pen-Type Pseudo-Haptic Interface Using Multipoint Suction Pressures," in SIGGRAPH Asia 2015 Haptic Media And Contents Design (New York, NY, USA: Association for Computing Machinery). doi:10.1145/2818384.2818399

Korres, G., and Eid, M. (2020). "Katib: Haptic-Visual Guidance for Handwriting," in International Conference on Human Haptic Sensing and Touch Enabled Computer Applications (Springer), 279–287. doi:10.1007/978-3-030-58147-3_31

Kyung, K.-U., Lee, J.-Y., and Park, J. (2007). "Design and Applications of a Pen-Like Haptic Interface With Texture and Vibrotactile Display," in 2007 Frontiers in the Convergence of Bioscience and Information Technologies, 543–548. doi:10.1109/fbit.2007.92

LaViola, J. J., and Zeleznik, R. C. (2004). "MathPad2," in ACM SIGGRAPH 2004 Papers (New York, NY, USA: Association for Computing Machinery), 432–440. doi:10.1145/1186562.1015741

Le Cornu, A. (2009). Meaning, Internalization, and Externalization. *Adult Education Q.* 59, 279–297. doi:10.1177/0741713609331478

Lin, L.-F., Teng, S.-Y., Liang, R.-H., and Chen, B.-Y. (2016). "Stylus Assistant," in SIGGRAPH ASIA 2016 Emerging Technologies (New York, NY, USA: ACM), 14:1–14:2. doi:10.1145/2988240.2988255

Lopes, P., Yuksel, D., Guimbretiere, F., and Baudisch, P. (2016). "Muscle-Plotter: an Interactive System Based on Electrical Muscle Stimulation that Produces Spatial Output," in 29th Annual Symposium on User Interface Software and Technology (New York, New York, USA: ACM Press), 207–217.

Magana, A. J., and Balachandran, S. (2017). Unpacking Students' Conceptualizations Through Haptic Feedback. *J. Computer Assist. Learn.* 33, 513–531. doi:10.1111/jcal.12198

Miller, B. E., Colgate, J. E., and Freeman, R. A. (2004). On the Role of Dissipation in Haptic Systems. *IEEE Trans. Robot.* 20, 768–771. doi:10.1109/tro.2004.829452

Minaker, G., Schneider, O., Davis, R., and MacLean, K. E. (2016). Handson: Enabling Embodied, Creative Stem E-Learning With Programming-Free Force Feedback. *Eurohaptics.*. Cham: Springer, 9775, 427–437. doi:10.1007/978-3-319-42324-1_42

Minogue, J., and Jones, G. (2009). Measuring the Impact of Haptic Feedback Using the Solo Taxonomy. *Int. J. Sci. Education.* 31, 1359–1378. doi:10.1080/09500690801992862

Montessori, M., and Carter, B. (1936). *The Secret of Childhood*. Calcutta: Orient Longmans Calcutta.

Mullins, J., Mawson, C., and Nahavandi, S. (2005). "Haptic Handwriting Aid for Training and Rehabilitation," in 2005 IEEE International Conference on Systems, Man and Cybernetics, 2690–2694. doi:10.1109/ICSMC.2005.1571556

Murayama, J., Bougrila, L., Luo, Y., Akahane, K., Hasegawa, S., Hirsbrunner, B., et al. (2004). Spidar G&G: a Two-Handed Haptic Interface for Bimanual Vr Interaction. *Proc. EuroHaptics (Citeseer).* 2004, 138–146.

Nakagaki, K., and Kakehi, Y. (2014). "Comp*Pass," in CHI '14 Extended Abstracts on Human Factors in Computing Systems (New York, NY, USA: ACM), 447–450. doi:10.1145/2559206.2574766

Özgür, A., Johal, W., Mondada, F., and Dillenbourg, P. (2017a). "Haptic-Enabled Handheld Mobile Robots," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (New York, NY, USA: ACM), 2449–2461. doi:10.1145/3025453.3025994

Özgür, A., Johal, W., Mondada, F., and Dillenbourg, P. (2017b). "Windfield: Demonstrating Wind Meteorology With Handheld Haptic Robots," in Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (New York, NY, USA: ACM), 48–49. doi:10.1145/3029798.3036664

Özgür, A., Lemaignan, S., Johal, W., Beltran, M., Briod, M., Pereyre, L., Mondada, F., and Dillenbourg, P. (2017c). "Cellulo," in ACM/IEEE Int'l Conf on Human-Robot Interaction, Vienna, Austria (IEEE), 119–127. doi:10.1145/2909824.3020247

Paczkowski, P., Kim, M. H., Morvan, Y., Dorsey, J., Rushmeier, H., and O'Sullivan, C. (2011). Insitu. *ACM Trans. Graph.* 30, 1–10. doi:10.1145/2070781.2024216

Papert, S. (1980). *Mindstorms: Children, Computers, and Powerful Ideas*. New York: Basic Books.

Pavlidis, T., and Van Wyk, C. J. (1985). "An Automatic Beautifier for Drawings and Illustrations," in Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (New York, NY, USA: Association for Computing Machinery), 225–234. doi:10.1145/325334.325240

Paynter, H. M. (1960). *Analysis and Design of Engineering Systems*. Cambridge, MA: MIT Press.

Perkins, K., Adams, W., Dubson, M., Finkelstein, N., Reid, S., Wieman, C., et al. (2006). Phet: Interactive Simulations for Teaching and Learning Physics. *Phys. Teach.* 44, 18–23. doi:10.1119/1.2150754

Perlin, K., He, Z., and Rosenberg, K. (2018). Chalktalk: A Visualization and Communication Language–As a Tool in the Domain of Computer Science Education *CoRR*. arXiv preprint arXiv:1809.07166

Piaget, J. (1964). Part I: Cognitive Development in Children: Piaget Development and Learning. *J. Res. Sci. Teach.* 2, 176–186. doi:10.1002/tea.3660020306

Piaget, J. (1977). *The Development of Thought: Equilibration of Cognitive structures.(Trans A. Rosin). (Viking)*. Washington, DC: Viking.

Radu, I., Hv, V., and Schneider, B. (2021). Unequal Impacts of Augmented Reality on Learning and Collaboration During Robot Programming With Peers. *Proc. ACM Hum.-Comput. Interact.* 4, 1–23. doi:10.1145/3432944

Reiner, M. (2009). Sensory Cues, Visualization and Physics Learning. *Int. J. Sci. Education.* 31, 343–364. doi:10.1080/09500690802595789

Renken, M. D., and Nunez, N. (2013). Computer Simulations and Clear Observations Do Not Guarantee Conceptual Understanding. *Learn. Instruction.* 23, 10–23. doi:10.1016/j.learninstruc.2012.08.006

Resnick, M., Martin, F., Berg, R., Borovoy, R., Colella, V., Kramer, K., et al. (1998). "Digital Manipulatives: New Toys to Think with," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Los Angeles, California, USA New York, NY, USA: ACM Press/Addison-Wesley Publishing Co), 281–287.

Resnick, M. (1998). Technologies for Lifelong Kindergarten. *ETR&D.* 46, 43–55. doi:10.1007/bf02299672

Romat, H., Henry Riche, N., Hinckley, K., Lee, B., Appert, C., Pietriga, E., and Collins, C. (2019). "ActiveInk," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (New York, NY, USA: Association for Computing Machinery), 1–13. doi:10.1145/3290605.3300272

Salisbury, J. K., and Srinivasan, M. A. (1997). Phantom-Based Haptic Interaction With Virtual Objects. *IEEE Comput. Grap. Appl.* 17, 6–10. doi:10.1109/mcg.1997.1626171

Seifi, H., Park, G., Fazlollahi, F., Sastrillo, J., Ip, J., Agrawal, A., et al. (2019). "Haptipedia: Accelerating Haptic Device Discovery to Support Interaction Engineering Design," in Conference on Human Factors in Computing Systems (CHI'19) (Glasgow, UK: ACM), 12.

Shannon, C. E. (1949). Communication in the Presence of Noise. *Proc. IRE.* 37, 10–21. doi:10.1109/jrproc.1949.232969

Smits-Engelsman, B. C. M., Niemeijer, A. S., and van Galen, G. P. (2001). Fine Motor Deficiencies in Children Diagnosed as Dcd Based on Poor Grapho-Motor Ability. *Hum. Movement Sci.* 20, 161–182. doi:10.1016/S0167-9457(01)00033-1

Teranishi, A., Korres, G., Park, W., and Eid, M. (2018). Combining Full and Partial Haptic Guidance Improves Handwriting Skills Development. *IEEE Trans. Haptics.* 11, 509–517. doi:10.1109/TOH.2018.2851511

Triana, M. J. R., Prieto, L. P., Vozniuk, A., Boroujeni, M. S., Schwendimann, B. A., Holzer, A., et al. (2017). Monitoring, Awareness and Reflection in Blended Technology Enhanced Learning: a Systematic Review. *Ijtel.* 9, 126–150. doi:10.1504/ijtel.2017.084489

Tsetserukou, D., Sato, K., and Tachi, S. (2010). "Exointerfaces: Novel Exosceleton Haptic Interfaces for Virtual Reality, Augmented Sport and Rehabilitation," in Proceedings of the 1st Augmented Human International Conference, 1–6.

Ullman, D. G., Wood, S., and Craig, D. (1990). The Importance of Drawing in the Mechanical Design Process. *Comput. Graphics.* 14, 263–274. doi:10.1016/0097-8493(90)90037-X

Vaquero-Melchor, D., and Bernardos, A. M. (2019). Enhancing Interaction With Augmented Reality Through Mid-air Haptic Feedback: Architecture Design and User Feedback. *Appl. Sci.* 9, 5123. doi:10.3390/app9235123

Wall, S. A., and Brewster, S. A. (2003). *Assessing Haptic Properties for Data Representation*. New York, NY, USA: Association for Computing Machinery, 858–859.

Wang, D., Zhang, Y., and Yao, C. (2006). Stroke-Based Modeling and Haptic Skill Display for Chinese Calligraphy Simulation System. *Virtual Reality.* 9, 118–132. doi:10.1007/s10055-005-0012-4

Weiser, M. (1991). The Computer for the 21st Century. *Sci. Am.* 265, 94–104. doi:10.1038/scientificamerican0991-94

Wobbrock, J. O., Wilson, A. D., and Li, Y. (2007). "Gestures Without Libraries, Toolkits or Training: A $1 Recognizer for User Interface Prototypes," in Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (New York, NY, USA: Association for Computing Machinery), 159–168. doi:10.1145/1294211.1294238

Yamaoka, J., and Kakehi, Y. (2013a). "Depend: Augmented Handwriting System Using Ferromagnetism of a Ballpoint Pen," in Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology (New York, NY, USA: Association for Computing Machinery), 203–210.

Yamaoka, J., and Kakehi, Y. (2013b). "dePENd," in ACM Symp on User Interface Software and Technology (UIST), 203–210. doi:10.1145/2501988.2502017

Young-Seok, K., Collins, M., Bulmer, W., Sharma, S., and Mayrose, J. (2013). "Haptics Assisted Training (HAT) System for Children's Handwriting," in 2013 World Haptics Conference (WHC), 559–564. doi:10.1109/WHC.2013.6548469

Zacharia, Z. C., Loizou, E., and Papaevripidou, M. (2012). Is Physicality an Important Aspect of Learning Through Science Experimentation Among Kindergarten Students?. *Early Child. Res. Q.* 27, 447–457. doi:10.1016/j.ecresq.2012.02.004

Zacharia, Z. C., and Michael, M. (2016). "Using Physical and Virtual Manipulatives to Improve Primary School Students' Understanding of Concepts of Electric Circuits," in *New Developments in Science and Technology Education* (Springer), 125–140. doi:10.1007/978-3-319-22933-1_12

Zeleznik, R., Bragdon, A., Adeputra, F., and Ko, H.-S. (2010). "Hands-on Math A Page-Based Multi-Touch and pen Desktop for Technical Work and Problem solving," in Proceedings of the 23nd annual ACM symposium on User interface software and technology, 17–26. doi:10.1145/1866029.1866035

Zinn, M., Khatib, O., Roth, B., and Salisbury, J. K. (2008). "Large Workspace Haptic Devices - a New Actuation Approach," in 2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, 185–192. doi:10.1109/haptics.2008.4479941

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 in Robotics and AI

ORIGINAL RESEARCH
published: 01 October 2021
doi: 10.3389/frobt.2021.699524

Personalizing HRI in Musical Instrument Practicing: The Influence of Robot Roles (Evaluative *Versus* Nonevaluative) on the Child's Motivation for Children in Different Learning Stages

Heqiu Song[1], Emilia I. Barakova[1], Panos Markopoulos[1] and Jaap Ham[2]*

[1]Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands, [2]Human-Technology Interaction, Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, Netherlands

OPEN ACCESS

Edited by:
Tony Belpaeme,
Ghent University, Belgium

Reviewed by:
Vicky Charisi,
Joint Research Centre, Belgium
Martina Blašková,
Police Academy of the Czech Republic, Czechia
Anara Sandygulova,
Nazarbayev University, Kazakhstan

***Correspondence:**
Emilia I. Barakova
e.i.barakova@tue.nl.

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: *23 April 2021*
Accepted: *19 August 2021*
Published: *01 October 2021*

Citation:
Song H, Barakova EI, Markopoulos P and Ham J (2021) Personalizing HRI in Musical Instrument Practicing: The Influence of Robot Roles (Evaluative Versus Nonevaluative) on the Child's Motivation for Children in Different Learning Stages. Front. Robot. AI 8:699524. doi: 10.3389/frobt.2021.699524

Learning to play a musical instrument involves skill learning and requires long-term practicing to reach expert levels. Research has already proven that the assistance of a robot can improve children's motivation and performance during practice. In an earlier study, we showed that the specific role (evaluative role versus nonevaluative role) the robot plays can determine children's motivation and performance. In the current study, we argue that the role of the robot has to be different for children in different learning stages (musical instrument expertise levels). Therefore, this study investigated whether children in different learning stages would have higher motivation when assisted by a robot in different supporting roles (i.e., evaluative role versus nonevaluative role). We conducted an empirical study in a real practice room of a music school with 31 children who were at different learning stages (i.e., beginners, developing players, and advanced players). In this study, every child practiced for three sessions: practicing alone, assisted by the evaluative robot, or assisted by the nonevaluative robot (in a random order). We measured motivation by using a questionnaire and analyzing video data. Results showed a significant interaction between condition (i.e., alone, evaluative robot, and nonevaluative robot) and learning stage groups indicating that children in different learning stage groups had different levels of motivation when practicing alone or with an evaluative or nonevaluative robot. More specifically, beginners had higher persistence when practicing with the nonevaluative robot, while advanced players expressed higher motivation after practicing with a robot than alone, but no difference was found between the two robot roles. Exploratory results also indicated that gender might have an interaction effect with the robot roles on child's motivation in music practice with social robots. This study offers more insight into the child-robot interaction and robot role design in musical instrument learning. Specifically, our findings shed light on personalization in HRI, that is, from adapting the role of the robot to the characteristics and the development level of the user.

Keywords: child–robot interaction, robots in music education, motivation for musical instrument practicing, robots for personalized education, robot roles, learning stages

1 INTRODUCTION

Musical instrument learning appears to have collateral cognitive benefits (Hassler et al., 1985; Anvari et al., 2002; Hietolahti-Ansten and Kalliopuska, 1990). Compared with passively listening to music, the strongest benefits were reported to come from active music listening and music training (Rauscher and Hinton, 2006). For example, singing, learning to play a musical instrument, and recognizing and keeping pitches and beat could improve learner's cognitive functions better than passive listening (Bernhard, 2002; Jausovec and Pahor, 2017). Due to the acquisition of a complex set of motor, sensory, and cognitive skills that learning a musical instrument requires (Lehmann et al., 2007) for beginners, it typically takes years to become skilled performers. Meanwhile, Ericsson et al. (1993) argue that the amount of deliberate practice is the major distinction between successful and unsuccessful learners through the long-term instrument learning process to reach a high-level achievement (Lehmann and Ericsson, 1997). Practicing at home is an important part of the instrument learning process for children (Hallam, 1998),but it is not an enjoyable activity for most children. It is crucial for teachers and parents to understand the significance of motivation in instrument learning, which is also a skill development process (Woody, 2004).

Besides practicing repeatedly, social factors also play an important role in the success of young children's music lessons. Previous literature indicates that young children depend dominantly on extrinsic motivation (Zdzinski, 1996), and, for example, peer support and peer tutoring can motivate the learner to engage in practicing music vigorously (Burnard, 2002), parental involvement is also a key factor that influences young children's persistence and achievement in instrument learning (Creech and Hallam, 2003; Moore et al., 2003). With the rapid development of social robots and the benefits of using robots in different educational contexts (Brown and Howard, 2014; Shiomi et al., 2015; Kennedy et al., 2016), social robots could also be used to provide social support to motivate children in instrument learning. And the main aim of using social robots in education is mainly to arouse children's motivation and improve the outcome of learning.

Earlier research suggested that the presence of the robot influences children's motivation and performance for learning activities (Riether et al., 2012). This can be explained by social facilitation theory (Triplett, 1898). More specifically, social facilitation theory (confirmed by earlier research, see, e.g., Zajonc (1965)) describes that people perform better at well-trained tasks when audience is present, while on new and complex tasks, people perform worse when they know they are being observed by other people (and robots) (Riether et al., 2012). In addition, several studies suggested that evaluation apprehension may be a necessary condition for producing social facilitation effects (e.g., Cottrell et al. (1968); Jones and Gerard (1967)), Sasfy also suggested that social facilitation only works under the condition of people trusting the audience to have the potential to evaluate (Sasfy and Okun, 1974). So it might not simply be the presence of a robot, but rather the role that it has in the social interaction that determines a user's response to the robot. That is, when a robot is present in an interaction as an evaluator of the user's behavior, the user's performance might be influenced. In general, earlier research has already deployed the role of robots as a tutor (Kennedy et al., 2016), a peer (Balkibekov et al., 2016; Park et al., 2017), a learner (Sandygulova et al., 2020; Hood et al., 2015), or a mediator of the interaction (Barakova et al., 2015).

Confirming the importance of the robot's role in stimulating music practice, in an earlier study (Song et al., 2020), we designed two robot roles (i.e., evaluative role and nonevaluative role) as companions for children's music practice and tested the influence of practicing while being assisted by a robot with one of the two roles on children's motivation and performance (Song et al., 2021). Results showed that children were more motivated and performed better with the nonevaluative robot.

However, as described above, social facilitation theory (Triplett, 1898) shows that for new and complex tasks, being observed by an evaluator might be detrimental for performance and motivation, but for well-trained tasks, being observed by an evaluator has positive effects on performance and motivation.

Especially in the domain of instrument learning, such evaluation seems important. That is, in the domain of instrument learning, a person's self-concept (a person's perception of themselves, see, Greenberg (1970)) may be more crucial than in other domains (Marsh et al., 2003). A person's self-concept is generated from an early age and more personalized when the person grows older (Marsh et al., 1991), which indicates that a learner's self-concept differs in different learning stages. And this perception can become more positive because of the evaluations from others or comparison with others (Greenberg, 1970; Bong and Clark, 1999; Lamont, 2011).

Still, in the domain of human-robot interaction, the effect of how evaluative a robot role is for children in different learning stages has not been established yet. Therefore, in the current study, we investigated whether children in different learning stages would have higher motivation when assisted by a robot in a different supporting role (i.e., evaluative role versus nonevaluative role). We conducted an empirical study in a real practice room of a music school with 31 children who were at different learning stages (i.e., beginners, developing players, and advanced players). In this study, every child practiced for three sessions, practicing alone, assisted by the evaluative robot, or assisted by the nonevaluative robot (in random order). We measured motivation by using a questionnaire and analyzing video data. We expected that children in different learning stages would have higher motivation with different robot roles (i.e., evaluative role versus nonevaluative role).

2 RELATED WORK

2.1 Musical Instrument Learning and Self-Regulation

From 1920s, researchers already started investigating the nonmusical benefits of musical training (Earhart, 1920; Hassler et al., 1985; Hietolahti-Ansten and Kalliopuska, 1990; Anvari et al., 2002). Additionally, as a skill learning activity, instrument learning requires amounts of practice through a combination of sensory input and output. The quantity and quality of practice have been one of the most important focuses in instrument learning, as well as collaborative music performance; evidence showed that it takes over 10 years for experts to train their skills to the level of a master (Ericsson et al., 1993; Williamon and Valentine, 2000; Thresher, 1964). Lehmann and Ericsson (1997) also pointed out, compared to unsuccessful learners, successful musicians put more effort into deliberate practice, which is one of the main reasons they reached a higher level. Practice has always played an important role in children's instrument learning; however, it is not enjoyable especially for young children (Hallam, 1998). In this case, parental involvement becomes quite important in children's persistence and achievement in instrument learning (Creech and Hallam, 2003; Moore et al., 2003). Nevertheless, with the increase of practice time, involvement of parents could exert pressure on children. And with time, it could eventually bring tense to the relationship between parents and their children.

In the development of children's instrument learning, their practice habits in different learning stages are varied; proper habits take years to develop. The result from a study by McPherson and Renwick (2001) showed that playing a music piece from the beginning to the end is one of the practicing strategies that most of the beginners applied. Parncutt and McPherson (2002) considered this as the lack of awareness of the mistakes they make, since beginners normally do not have the ability to identify and correct their own mistakes. After investigating the practice habits of musicians and learners, researchers have also paid attention to the myriad factors implicated in musical learning, especially in musical instrument practicing, which offered a crucial backdrop for self-regulation research in instrument learning context. Started with McPherson and his colleagues from the 1990s, researchers started to investigate the educational construct which was known as self-regulation. According to Zimmerman's sociocognitive model of self-regulation, personal perceptions, efficacy, and environmental conditions are involved in self-regulated learning (SRL) (Zimmerman, 1989), which indicated that a social context or an environment is an important part of students' SRL. However, as a self-regulated activity, few researchers have applied self-regulation for social robot design in musical instrument learning. And self-regulated learning is related to the expertise level (Varela et al., 2016); we chose to focus on children's learning stages in this study.

2.2 Motivation in Instrument Learning

Since social support is a key component in self-regulated learning and practicing is an important but not enjoyable part of the instrument learning process for children (Hallam, 1998), keeping children motivated during instrument learning seems important for teachers and parents. Self-determination theory has mainly been used to explain motivation in instrument learning (Renwick and Reeve, 2012; MacIntyre et al., 2018). This theory proposes that people have three core psychological needs, which are autonomy, competence, and relatedness. These psychological needs will be satisfied to different levels (Deci and Ryan, 2000; Ryan and Deci, 2000). Furthermore, as a crucial factor that can influence motivation to participate in learning activities, a student's self-concept has found to be stronger in music than in other domains (Marsh et al., 2003), which means in order to maintain the motivation to persist in instrument learning, a strong music self-concept is a crucial component for it. However, as children age, these beliefs tend to change. Younger students are more inclined to have positive achievement beliefs. As they grow older, they tend to become more realistic about how successful they will be (Wigfield et al., 1997). This could be explained by children's development of self-concept and the evaluation or feedback from others (Ireson and Hallam, 2009). Hence, it is intriguing to study the influence of evaluation on learners in different learning stages with different levels of self-concept.

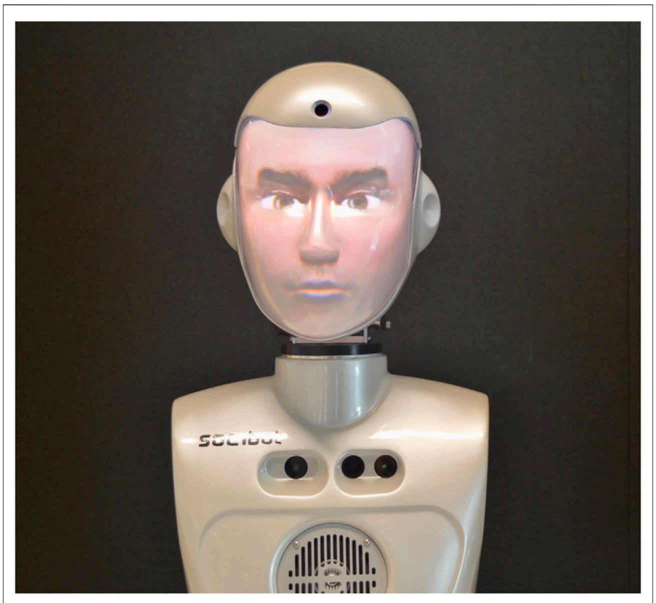


FIGURE 1 | The SocibotMini robot used in the current study.

2.3 Social Robot for Educational Use

As promising tools in education, educational robots are becoming a popular research topic, since there are proof that showed using a robot in the educational context could improve teachers' effectiveness and students' learning motivation. (Johnson et al., 2003; Papert, 1993). With the rapid development of technology and social robotics research, a lot of challenges researchers faced had already been solved or can be taken care of soon. For instance, in instrument learning, with the combination of sensors and cameras in a robot or other devices, the robot should be able to detect and correct wrong postures. There is also a well-developed system that we can build in the robot to detect the mistakes and evaluate the pitch and tempo of the performance (Asahi et al., 2018). We believe that student's musicality can be developed well with more accurate judgements and individual strategies from the robot. We also envisioned the role of robot in children's instrumental learning mainly as a companion, which is able to provide professional help and social support in children's daily practice with the capability of 24/7 availability, flawless memory, mistake detection, big data, professional music knowledge, different teaching strategies, and so on. In the future, the robot could also be used to provide music teachers help in the music lessons to offer each child suitable learning experience. Research has found that interacting with a tangible robot resulted in more engagement than interacting with a video (Xie et al., 2008). Various educational scenarios have been employed with a social robot, including knowledge learning [e.g., math (Brown and Howard, 2014), science (Shiomi et al., 2015)] and skills learning [e.g., music (Han et al., 2009) and language (Nielsen et al., 2008)].

In educational scenarios, the roles of robots have also been explored. The robot can be deployed as a tutor (Kennedy et al., 2016), a peer (Balkibekov et al., 2016; Park et al., 2017), a learner (Sandygulova et al., 2020; Hood et al., 2015), or a mediator of the interaction (Barakova et al., 2015). By implementing the social facilitation theory and the evaluation apprehension theory, in the previous study (Song et al., 2020), we developed two roles (i.e., evaluative role and nonevaluative role) in the context of musical instrument practice by using SocibotMini (see **Figure 1**), which is also the robot we used in the current study, a robot with a projected face that provides rich human-like expressions. In the previous study, we performed the evaluation of those two roles in a real practice room at the music school. The study concluded that the designed evaluative and nonevaluative role of the social robot was convincing and matched children's cognitive expectations in the music practicing context (Song et al., 2020).

2.4 Research Questions and Hypotheses

To figure out the impact of robot roles (i.e., evaluative role and nonevaluative role) on the motivation of children in different learning stages, we proposed the following:

- Research Question: Can different robot roles (i.e., evaluative role and nonevaluative role) affect the motivation of children in different learning stages differently in instrument practice?
- Hypothesis: We expect to find an interaction between the learning stage and robot condition (i.e., alone vs evaluative role vs nonevaluative role) on children's motivation in instrument practice.

Furthermore, we will explore whether other factors influence the impact of different robot roles on children's motivation in instrument practice.

3 METHOD

3.1 Research Context and Participants

This study was conducted at a common piano practicing room of a music school (Centrum voor de Kunsten Eindhoven, CKE) in the Netherlands. We chose the piano because the piano is one of the most popular musical instruments with children. In addition, it is quite easily possible to match the difficulty level of a new melody with the music learning level of a child, and also because of practical reasons (e.g., number of available potential participants at the local music school).

Two rooms were used for the experiment: an experiment room (see **Figure 2A**) for the participants to practice and interact with the robot and a control room (see **Figure 2B**) for the researcher to observe the process and control the robot. Two cameras (camera A and camera B) were set up in the experiment room before the experiment started. In the control room (see **Figure 2B**), researchers and parents could watch the child and the robot in the experiment room via the monitor which showed the view of camera B. We used the Wizard of Oz method to ensure situation-specific utterances of the robot. Researcher A in the control room monitored the experiment on laptop A (with the view of camera A in the practicing room) and the monitor (showing the view of camera B in the practicing room) to control the robot through laptop B. Camera A recorded the whole practicing session of each child.

The participants were 31 children ($N = 31$) aged from nine to 12 years old who were taking piano lessons in the music school. The learning stage of children, which indicates their level of expertise, ranged from 2 months to 5.5 years. Children were divided into three different learning stage groups based on the suggestion from piano teachers, which are beginners (had learned piano for less than 2 years, $n = 11$), developing players (had learned piano for two to 4 years, $n = 10$), and advanced players (had learned piano for more than 4 years, $n = 10$). The number of girls ($n = 15$) was approximately equal to that of boys ($n = 16$).

3.2 Robot Roles

We employed the two robot roles that we designed and evaluated in the previous study (Song et al., 2020). The evaluative role of the current study used "forceful and concrete language, providing praise on effort, with a slow and steady pitch, and a calm facial expression, focusing on the practice, move little, and dress formally (i.e., shirt)" (Song et al., 2020). In contrast, the robot in the nonevaluative role used "indirect and abstract language, provided praise on talent, with a quick and active pitch, and a funny facial expression, moved a lot, and dressed informally (i.e., striped sweater)" (Song et al., 2020). In the previous study, the interaction scripts for each role of the robot included 36 behaviors (Song et al., 2020). In the current study, children needed to practice two melodies in each session. To measure the performance data correctly, it is better if children can keep practicing one melody first and then change to the other one. In this case, we designed six new behaviors for the robot to indicate the melody the children should practice during practicing sessions. We also deleted some behaviors that were proved not suitable in the practicing context to simplify the interaction, but still kept the robot roles clear enough. In total, there were 30 behaviors of the robot adapted and designed (see **Table 1**).

In addition, for the purpose of eliminating the 'novel effect' of practicing with a social robot, each child went through an introduction session with a robot, which has a neutral role. Twelve behaviors of the robot (i.e., a robot introduces itself

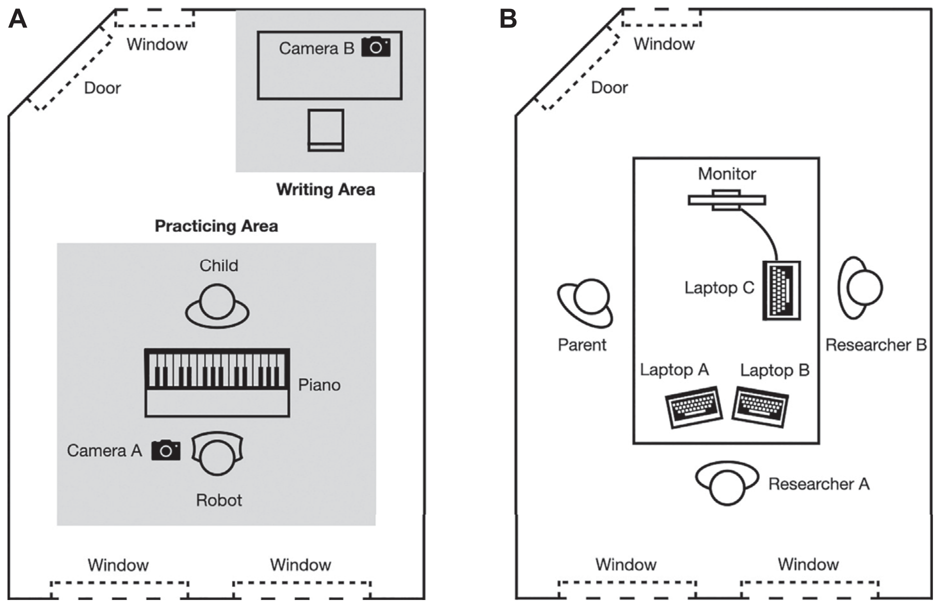


FIGURE 2 | Top view illustration of the experiment room **(A)** and the control room **(B)**.

TABLE 1 | Numbers of robot behaviors for evaluative role and nonevaluative role in different tasks.

Context	Tasks	Evaluative robot	Nonevaluative robot
Robot introduction	Robot self-introduction	1	1
	Greeting	2	2
	General task introduction	3	3
Practice Session	General task introduction	1	1
	Melody order guide	2	2
	Verbal feedback for praise	3	3
	Verbal feedback for stop	1	1
	Verbal feedback for playing wrong melody	1	1
	Verbal feedback for questions	3	3
	Conclusion of the practice	1	1
Filling in questionnaire	Ask to fill in	2	2
	Ask to take a break	1	1

and introduces the practicing task generally) were employed here in the introduction session to generate a fluent conversation (see **Table 1**).

3.3 Measures

We adapted and combined three subscales (i.e., autonomy, delight, and stress) from the FunQ questionnaire (Tisza and Markopoulos, 2021) and three questions from The Situational Motivation Scale (SIMS; Guay et al. (2000)) to measure the children's motivation (see **Table 2**). The combined questionnaire consisted of four dimensions of motivation: autonomy, delight, stress, and interest. The FunQ questionnaire was developed to measure the fun value of a learning activity with adolescents around 12 year old, which is positively correlated with engagement (Iten and Petko, 2016; Rambli et al., 2013). We took the autonomy, delight, and stress dimensions from it as indicators of motivation. As for the interest dimension of the questionnaire, the questions in this dimension were adapted from the SIMS, in which questions were not included in the three dimensions mentioned before. Since the SIMS was not designed for children, based on the findings of Mellor and Moore (Mellor and Moore, 2014), the questions were changed into 5-point Likert questions and made easier by simplifying the language to the level of the children. Each dimension had three questions. Some questions had to be reverse coded: one question for the autonomy dimension and all three questions in the stress dimension were reversed questions that needed to be analyzed oppositely (**Table 2**). The motivation was also measured with observation data from the videos from camera A in **Figure 2A**.

3.4 Procedure

The participants were invited to the study through emails, and teachers from CKE helped us by sending the emails to the parents.

TABLE 2 | Motivation questionnaire questions.

Source	Dimension	Question
FunQ	Autonomy	I knew what to do
	—	I did this activity because I had to. (r)
	—	I did this activity because I wanted to
FunQ	Delight	I was happy
	—	I had fun
	—	I want to do something like this again
FunQ	Stress	I felt angry. (r)
	—	I felt sad. (r)
	—	I felt bad. (r)
SIMS	Interest	I could focus easily
	—	I think this practice is important
	—	I did this activity because I wanted to

r: reversed items.

After the participants and their parents arrived at the assigned room in the music school, they were asked to sign the informed consent form of the current study and fill in the questionnaire to collect basic information (e.g., gender, age, and learning duration). Then, researcher B explained the procedure in detail to them and escorted participants and their parents into the experiment room (see **Figure 2A**). In the meantime, the robot started to greet the children and encouraged them to talk to them. Next, the parents were invited to the control room while every child was practicing piano in three conditions (i.e., alone, with the evaluative robot, and with the nonevaluative robot) in random order, in a within-subject experiment design. In each condition, children needed to practice for 10 minutes (5 minutes for each music piece). At the beginning of the robot conditions, the robot gave a self-introduction first, including the name (i.e., "Jimmy" is the evaluative robot and "Peter" is the nonevaluative robot). After each session, they were asked to fill in the questionnaire to measure their motivation. And between each session, children were allowed to take a 5-minute break.

The melodies we chose to use in the experiment are parts of Chinese children's songs because we want to make sure these are totally new to the (Dutch) participants. Eventually, we had in total nine pieces for each of the learning stage groups (i.e., beginners, developing players, and advanced players) and conditions (i.e., alone, evaluative robot, and nonevaluative robot). For instance, piece one is for beginners to play in the alone condition, piece two is for beginners to play in the evaluative robot condition, and piece four is for developing players to play in the alone condition. All the pieces were selected by a music teacher who has more than 25 years of experience of teaching the piano. The difficulty levels of the selected melodies were super easy for beginners (can be played by a single hand), easy for developing players (starting level of add another hand, for instance, at the start of each bar), and medium for advanced players (played by two hands, multiple keys at the same time). The duration of all the pieces was around 15 s since we only offered 5 min for them to practice each piece.

Regarding to the complexity of children's motivation in instrument learning, our main purpose of this study is to test the different effects of different robot roles on children's motivation. During the whole experiment, we kept all the possible confounding variables the same in the three conditions. The only difference between conditions is the type of accompanying (i.e., none (alone), evaluative robot, and nonevaluative robot) we provided. And thereby, we control for the influence of any other variable.

3.5 Data Analysis

The questionnaire data and video data were used to address whether children in different learning stages would have higher motivation with robots in different supporting roles (i.e., evaluative role versus nonevaluative role) in musical instrument practicing. First of all, by averaging the answers from the twelve questions, we were able to construct reliable measures for each dimension of the motivation scale, that is, autonomy (*Cronbach's alpha* = 0.65), delight (*Cronbach's alpha* = 0.80), stress (*Cronbach's alpha* = 0.73), interest (*Cronbach's alpha* = 0.77), and the questionnaire in total (*Cronbach's alpha* = 0.89). To ascertain the impact of different roles of the robot on children's motivation in practice, we divided the children into three different learning stage groups based on the suggestion from piano teachers. Additionally, as a nonempirical method, length of experience has been used as one of the factors to identify expertise in the clinic context (Rassafiani, 2009). The learning stage groups are beginners (had learned piano for less than 2 years), developing players (had learned piano for two to 4 years), and advanced players (had learned piano for more than 4 years). After that, we performed a repeated measure ANOVA (see Results section). Additionally, to increase the reliability and interpretability of the measurement for children's motivation, despite the questionnaire, the video of children we took during the practicing was also analyzed. Because our video recordings (taken by camera A in **Figure 2A**) of the task were primarily focused on children's faces and upper bodies, we coded the video data by using a validated coding scheme for assessing motivation in young children developed by Berhenke et al. (2011). They explored the importance of emotions in young children's motivation, which they coded through facial movements (Izard, 1971). Meanwhile, they also took mastery motivation and strategy use as observed indicators of motivation. Eventually, they developed a coding system including four categories of codes, which were emotion states (neutral, positive, interest/arousal, sadness, confusion, anxiety, and anger), emotion events (pride, frustration, hostility, and shame), task behavior states (persistence, socializing and off-task), and task behavior events (help-seeking and competence). We adapted their coding system to fit the context of children's piano practicing with the social robot by deleting codes that are not important in this context, adding behaviors that especially exist in musical instrument practicing, and adding codes to observe children's interaction with the robot (Perugia et al., 2018). Finally, we analyzed the video data with the codes listed in **Table 3**.

Emotional expressions, task-related behaviors, and robot-related behaviors were coded independently by two independent researchers using the Observer XT 15.0 software system (Zimmerman et al., 2009) (*Cohen's* $k = 0.673$, $p < 0.01$). Except for the persistence and off-task codes in the task-related

TABLE 3 | Descriptions for Emotion and Task Behavior Codes.

Category	Positive codes and indicators	Negative codes and indicators
Emotional expressions	**Happiness**	**Sadness**
	*Laughs/giggles	*Sad Expression
	*Grins/smiles	*Self-frustration
	*Pride	**Anxiety**
	—	*Anxious expression
	—	*Shame/gaze avoidance/face hidden
	—	*Confusion/frozen expression
Task-related behaviors	**Persistence**	**Hostility/Reluctance**
	-Visual focus point on task	*Stop playing
	*Showing initiative	*Lean away from piano/leave
	*Application of personal strategies	*Shortcuts
	—	*Frustrated verbal remarks
	Help-seeking	**Off-task**
	*Verbally, directly, and explicitly ask for help	*Signs of boredom
	—	-Visual focus off task
Robot-related behaviors	**Interest**	**Hostility/Reluctance**
	-Visual focus on robot	*Refuses/hesitates/ignore to follow robot's directions
	*Engaged with a robot	*Verbal remarks "must I really"
	*Curiosity toward robot	—
	*Reply to the robot positively	—
	*Follows instructions	—
	Help-seeking	**Socializing**
	*Chats with the robot about the task	*Chats about other topics but the task

**Dichotomous nominal code.*
**Continuous code.*

behaviors category, which was coded continuously and analyzed by the percentage of focusing time in each mutually task-related behavior, the rest codes in the coding system were dichotomous nominal codes, which were counted and those counts were used in the analysis. Facial, vocal, and behavioral cues were used to indicate emotional expressions (e.g., triangular brows for sadness and smile for happiness) and task-related behaviors (e.g., intently keep practicing). Afterward, we performed independent sample *t* test to examine the impact of robots on children's motivation and offer an additional explanation to the results of questionnaire data.

4 RESULTS

In order to investigate the impact of robots in different supporting roles (i.e., evaluative role versus nonevaluative role) on the motivation of children in different learning stages in musical instrument practicing, the following analyses were performed. For the questionnaire data, firstly, we tested the normality of all questions in the questionnaire by using the Kolmogorov–Smirnov test and Shapiro–Wilk test. Results indicated that the answers to each dimension and question follow a normal distribution ($p < 0.05$). Then, we performed repeated-measures ANOVA to examine the impact of robot roles on children's motivation (RQ). Results showed a main effect of the practicing condition (i.e., alone, with the evaluative robot and with the nonevaluative robot) on the children's motivation measured using the questionnaire (F (2, 29) = 4.14, p = 0.02). Pairwise comparisons between the conditions showed that children had higher motivation after interacting with the evaluative robot (M = 4.53, SD = 0.77) and with the nonevaluative robot (M = 4.55, SD = 0.37) than after practicing alone (M = 4.34, SD = 0.56), with both p's < 0.05.

Furthermore, confirming our first hypothesis, on *motivation as measured using the questionnaire*, results showed an interaction effect between robot role conditions (i.e., alone, with the evaluative robot, and with the nonevaluative robot) and learning stage groups [F (2, 29) = 2.88 and p = 0.03]. This indicated that children in different learning stage groups had different levels of motivation in the three conditions. More specifically, as shown in **Table 4**, compared to beginners (M = 4.40 and SD = 0.40) and developing players (M = 4.56 and SD = 0.39), advanced players showed the tendency to have the lowest motivation in the alone condition (M = 4.07 and SD = 0.97), as indicated by t = 1.69 and p < 0.10. For the two robot conditions (i.e., the evaluative role and nonevaluative role), such differences in motivation as measured by the questionnaire were not found, with both t's < 1.00.

Additionally, we explored the *behavioral measures of motivation* and found comparable results on one of the indicators of motivation: persistence, which is very clearly an indicator on the participant's motivation in instrument learning; see **Table 5**. For this, just as for the questionnaire data, we tested the normality of all the codes and the results showed that except for the interest in the 'robot-related behaviors' category, which is not calculated here in the comparison of motivation on task, all observational data follow a normal distribution (p < 0.05). On these data, we performed independent sample t tests. Confirming

TABLE 4 | Mean and standard deviation of motivation collected by the questionnaire in the alone, nonevaluative robot, and evaluative robot conditions.

	Alone			Nonevaluative robot			Evaluative robot		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Beginners	11	4.40	0.40	11	4.50	0.42	11	4.43	0.29
Developing players	10	4.56	0.39	10	4.63	0.33	10	4.53	0.43
Advanced players	10	4.07	0.97	10	4.53	0.39	10	4.65	0.56
Total	31	4.34	0.65	31	4.55	0.37	31	4.53	0.77

TABLE 5 | Mean and standard deviation of motivation measured as persistence collected by the behavior data in the alone, nonevaluative robot, and evaluative robot conditions. Videos with a bad quality were not analyzed (incomplete, too much noise, etc.).

	Alone			Nonevaluative robot			Evaluative robot		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Beginners	7	4.71	2.43	7	3.00	2.94	7	3.57	1.51
Developing players	9	3.67	3.57	9	2.11	2.57	9	2.33	3.24
Advanced players	10	2.80	2.39	10	1.90	1.60	10	0.90	1.10
Total	26	3.62	2.86	26	2.27	2.31	26	2.12	2.36

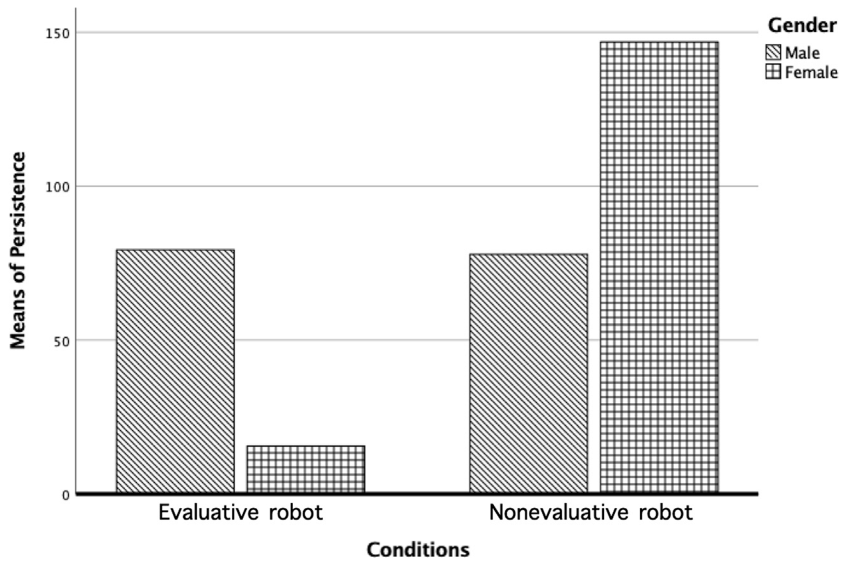


FIGURE 3 | Means of persistence of children in different gender groups in the evaluative robot condition and the nonevaluative robot condition.

our finding on *motivation as measured using the questionnaire*, these analyses showed that advanced players (M = 2.80 and SD = 2.39) had lower persistence than beginners (M = 4.71 and SD = 2.43) and developing players (M = 3.67 and SD = 3.57) in the alone condition [t (2, 29) = 2.12 and p = 0.04]. It also showed that beginners' persistence (M = 3.00 and SD = 2.94) was significantly higher than the advanced players (M = 1.90, SD = 1.60) in the nonevaluative condition [t (2, 29) = 4.03 and p = 0.00]. Furthermore, beginners also had higher persistence than developing players and advanced players, as indicated by a significant trend [t (2, 29) = 1.90 and p = 0.07]. For the evaluative robot conditions, such differences in motivation as measured by behavioral measures of motivation were not found, t < 1.00.

After we compared the difference between learning stage groups in each of the conditions, we performed a comparison within each of the learning stage groups as well. We conducted paired sample t test on *motivation as measured by the questionnaire* within learning stage groups, and the results showed that advanced players had higher motivation in both robot conditions (i.e., evaluative role (M = 4.65 and SD = 0.56) and nonevaluative role (M = 4.53 and SD = 0.39)) than the alone condition [M = 4.07, SD = 0.97, t (2, 29) = 3.39, and p = 0.01; t (2, 29) = 2.02 and p = 0.07]. For the other two learning stage groups (i.e., beginners and developing players), such differences in motivation as measured by the questionnaire were not found, with both t's < 1.00.

Furthermore, to explore other factors that may affect children's motivation in music practicing with different roles of the robot, we also analyzed the difference between genders. According to the result we got from MANOVA, gender might also be a factor that influences children's motivation through persistence in music practice [F (1, 30) = 5.02, p = 0.03]. As shown in **Figure 3**, under the nonevaluative robot condition, girls tended to have better persistence than boys, which means they focus on practice longer and follow more instructions. On the opposite, girls tended to persist less than boys in the evaluative robot condition while boys' persistence level remained similar in both of the conditions. Except gender, we also checked whether age is the factor that influences children's preference of the robot roles. Therefore, we conducted a correlation between the learning duration and age. Finally, the result showed no significant correlation between the learning stage and age (r = -0.01, p = 0.96), which suggested that the age of children did not influence the results we found on children's learning stages. Also, this indicates that although younger people might have liked more the nonevaluative robot, younger people were not necessarily beginners.

5 DISCUSSION AND CONCLUSION

The aim of the current research was to find out the impact of robots in different supporting roles (i.e., evaluative role versus nonevaluative role) on children's motivation in different learning stages in musical instrument practice. First of all, we performed a repeated-measure ANOVA to answer the first research question (RQ): Can robot roles affect the motivation of children in different learning stages differently? The main effect firstly indicated that regardless of the learning stages children in, children tended to have higher motivation with the nonevaluative role. Then, the result of the interaction effect combined with the result from behavioral data showed differences between different learning groups, which could answer the research question. With the result we got from the questionnaire and video data, we were able to confirm our hypothesis and suggested that children in different learning stages are more motivated with robots with different roles during practicing. This finding is in line with self-concept development in instrument learning (Greenberg, 1970) and social facilitation theory (Zajonc, 1965). That is, beginners have not developed clear acknowledgment about their abilities, which may have a negative impact on their motivation; a positive audience might encourage them better than an evaluative judge. Additionally, beginners do not yet have the (evaluation) skills to estimate which evaluation information from others can be used. As for the advanced players, their music self-concepts have been greatly reinforced by earlier music experiences, which makes them conscious of their own ability (Ireson and Hallam, 2009). In other words, advanced players usually have a higher self-concept. Without companionship, it is easy for them to lose patience in instrument practice. The encouragement and feedback provided by the robot may reinforce the self-concept of the advanced players (Greenberg, 1970; Bong and Clark, 1999; Lamont, 2011), which further improved their motivation in practicing. Combined with our empirical results, we believe that applying different roles of the robot in different learning stages would help children keep motivated in music instrument learning. More specifically, we suggest employing social robots in children's instrument learning, especially for the advanced players. Also, it seems better to use less evaluative robot for the beginners who just started learning a musical instrument. At the same time, we suggest that individual characteristics of learners during piano practicing might require more complex pedagogical approaches.

In addition, results from the observations further strengthened our confidence to confirm the conclusion. Within all the indicators that we coded for motivation, the most important result is that robot roles have an impact on children's persistence in different learning stages. According to the definition of motivation as a "process whereby goal-directed activity is energized and sustained" (Pintrich and Schunk, 2002), persistence is a key indicator of sustainment. This result aligns with the questionnaire result, which showed advanced players had better persistence with a robot than alone and beginners tend to have better persistence with the nonevaluative robot.

While the main results were in accordance with our expectations, the exploratory results also suggested that gender might be a determinant of a user's evaluation of the robot, in interaction effect with the robot's role. A potential explanation might be that the interaction between gender and robot role influenced children's motivation through persistence in music practice. However, in the review about gender differences in self-concept with children and adolescents done by Wilgenbusch and Merrell (1999), female participants reported significantly higher levels of self-concept in the musical domain among elementary grade participants (grade 1–6). Applying this finding to our case, girls should persist longer with the evaluative robot, which contrasts with our result. Still, Wilgenbusch and Merrell (1999) also indicated that their results were based on quite small effect sizes. And they also found that males reported significantly stronger self-concepts in the musical domain among secondary grade participants (grade 7–12). Our participants consist of children in both of their grade groups, and therefore, further work needs to be carried out to establish whether gender is a crucial factor that can affect children's motivation in the child–robot interaction.

Although our findings indicated the idea that different roles of the robot should be employed in different stages of children's instrument learning, we are aware that our research may have two limitations and which future studies can extend. The first is about the appearance and function of the robot. We used the SocibotMini, which does not have arms and cannot move. These features limited the possibility of interactions and may cause distrust from children on robot's music-related abilities. Furthermore, this robot requires real-time operation from the controlling system to be able to react.

However, it was impossible for the researcher to eliminate the delay in the interaction between children and the robot, even with a real-time surveillance camera. Therefore, there should be an inevitable influence on the interaction. This can be improved by using a full-body robot with an intelligent system. Secondly, even though we investigated child–robot interaction design for long-term companions with different learning stage groups, another perspective is studying changes in learning styles within one individual. For example, with the children's grows, their perception and preference would change. It is important to investigate the link between children's age and robot role preference in a long-term interaction design. Furthermore, instead of using children in different learning stage groups, future experimental investigations are needed to find out children's perception and preference of the robot roles within the development of the same individual. By then, we may focus on discovering the long-term effect of social robots in music instrument learning. In future research, it is also crucial to investigate more detailed factors that can affect children's motivation and performance in instrument learning, for instance, find out the impact of different kinds of evaluation in the instrument learning context (Sasfy and Okun, 1974). Another interesting research direction to investigate would be music-induced emotions. Emotion experiences are valid indicators for motivation (Csikszentmihalyi, 2000; Pekrun, 2006) and emotion expressions (e.g., facial expressions) are measurable elements of people's emotion experience. In the context of instrument learning, some of the basic indicators (e.g., smile) are rather rare in instrument learning and practice. However, not showing a smile is not equal to not being happy, as the expressions of positive emotions may be hidden in the instrument learning context. In this case, it is valuable to investigate whether it is correct to use common indicators to measure motivation in instrument learning.

In conclusion, our study confirmed the impact of different robot roles in children's instrument learning process, which offers more insight into child–robot interaction and robot role design in musical instrument learning. Specifically, our findings shed light on personalization in HRI, that is, adapting the role of the robot to the characteristics and the development level of the user.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion 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 Ethics Review Board Eindhoven University of Technology. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

HS, EB, JH, and PM designed the experiments and contributed to the analysis of the result. HS carried out the experiments. HS wrote the manuscript in consultation with EB, JH, and PM.

FUNDING

This research was funded by the China Scholarship Council.

ACKNOWLEDGMENTS

The authors would like to show their gratitude to all the teachers, parents, and children involved in the research project.

REFERENCES

Anvari, S. H., Trainor, L. J., Woodside, J., and Levy, B. A. (2002). Relations Among Musical Skills, Phonological Processing, and Early reading Ability in Preschool Children. *J. Exp. Child Psychol.* 83, 111–130. doi:10.1016/s0022-0965(02)00124-8

Asahi, S., Tamura, S., Sugiyama, Y., and Hayamizu, S. (2018). "Toward a High Performance Piano Practice Support System for Beginners," in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (Honolulu, HI, USA: IEEE), 73–79. doi:10.23919/apsipa.2018.8659463

Balkibekov, K., Meiirbekov, S., Tazhigaliyeva, N., and Sandygulova, A. (2016). "Should Robots Win or Lose? Robot's Losing Playing Strategy Positively Affects Child Learning," in 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (New York, NY, USA: IEEE), 706–711. doi:10.1109/ROMAN.2016.7745196

Barakova, E. I., Bajracharya, P., Willemsen, M., Lourens, T., and Huskens, B. (2015). Long-term LEGO Therapy with Humanoid Robot for Children with ASD. *Expert Syst.* 32, 698–709. doi:10.1111/exsy.12098

Berhenke, A., Miller, A. L., Brown, E., Seifer, R., and Dickstein, S. (2011). Observed Emotional and Behavioral Indicators of Motivation Predict School Readiness in Head Start Graduates. *Early Child. Res. Q.* 26, 430–441. doi:10.1016/j.ecresq.2011.04.001

Bernhard, H. C. (2002). Singing in Instrumental Music Education: Research and Implications. *Update Appl. Res. Music Educ.* 22, 28–35. doi:10.1177/87551233020220010501

Bong, M., and Clark, R. E. (1999). Comparison between Self-Concept and Self-Efficacy in Academic Motivation Research. *Educ. Psychol.* 34, 139–153. doi:10.1207/s15326985ep3403_1

Brown, L. N., and Howard, A. M. (2014). "The Positive Effects of Verbal Encouragement in Mathematics Education Using a Social Robot," in 2014 IEEE Integrated STEM Education Conference (Princeton, NJ, USA: IEEE), 1–5. doi:10.1109/isecon.2014.6891009

Burnard, P. (2002). Investigating Children's Meaning-Making and the Emergence of Musical Interaction in Group Improvisation. *Br. J. Mus. Ed.* 19, 157–172. doi:10.1017/S0265051702000244

Cottrell, N. B., Wack, D. L., Sekerak, G. J., and Rittle, R. H. (1968). Social Facilitation of Dominant Responses by the Presence of an Audience and the Mere Presence of Others. *J. Personal. Soc. Psychol.* 9, 245–250. doi:10.1037/h0025902

Creech, A., and Hallam, S. (2003). Parent-teacher-pupil Interactions in Instrumental Music Tuition: a Literature Review. *Br. J. Music. Ed.* 20, 29–44. doi:10.1017/S0265051702005272

Csikszentmihalyi, M. (2000). *Beyond Boredom and Anxiety*. 25th anniversary ed edn. San Francisco: Jossey-Bass Publishers

Deci, E. L., and Ryan, R. M. (2000). The "What" and "Why" of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychol. Inq.* 11, 227–268. doi:10.1207/S15327965PLI1104_01

Earhart, W. (1920). *The Value of Applied Music as a School Subject*. Hartford: Music Teacher National Association, 163–170.

Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychol. Rev.* 100, 363–406. doi:10.1037/0033-295X.100.3.363

Greenberg, M. (1970). Musical Achievement and the Self-Concept. *J. Res. Music Educ.* 18, 57–64. doi:10.2307/3344358

Guay, F., Vallerand, R. J., and Blanchard, C. (2000). On the Assessment of Situational Intrinsic and Extrinsic Motivation: The Situational Motivation Scale (SIMS). *Motiv. Emot.* 24, 175–213. doi:10.1023/a:1005614228250

Hallam, S. (1998). The Predictors of Achievement and Dropout in Instrumental Tuition. *Psychol. Music* 26, 116–132. doi:10.1177/0305735698262002

Han, J.-H., Kim, D.-H., and Kim, J.-W. (2009). *Physical Learning Activities with a Teaching Assistant Robot in Elementary School Music Class*. Seoul: IEEE, 1406–1410.

Hassler, M., Birbaumer, N., and Feil, A. (1985). Musical Talent and Visual-Spatial Abilities: A Longitudinal Study. *Psychol. Music* 13, 99–113. doi:10.1177/0305735685132004

Hietolahti-Ansten, M., and Kalliopuska, M. (1990). Self-Esteem and Empathy Among Children Actively Involved in Music. *Percept Mot. Skills* 71, 1364–1366. doi:10.2466/pms.1990.71.3f.1364

Hood, D., Lemaignan, S., and Dillenbourg, P. (2015). "When Children Teach a Robot to Write: An Autonomous Teachable Humanoid Which Uses Simulated Handwriting," in Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (Portland Oregon USA: ACM), 83–90. doi:10.1145/2696454.2696479

H. W. Marsh, R. G. Craven, and D. M. McInerney (2003). in *International Advances in Self Research* (Greenwich, Conn: Information Age Pub)

Ireson, J., and Hallam, S. (2009). Academic Self-Concepts in Adolescence: Relations with Achievement and Ability Grouping in Schools. *Learn. Instr.* 19, 201–213. doi:10.1016/j.learninstruc.2008.04.001

Iten, N., and Petko, D. (2016). Learning with Serious Games: Is Fun Playing the Game a Predictor of Learning success?. *Br. J. Educ. Technol.* 47, 151–163. doi:10.1111/bjet.12226

Izard, C. E. (1971). English. *The Face of Emotion (New York: Appleton-Century-Crofts)*. OCLC: 161033

Jausovec, N., and Pahor, A. (2017). English. *Increasing Intelligence*. London: Academic Press. OCLC: 1062390226.

Johnson, C. A., Adams, J. A., and Kawamura, K. (2003). "Evaluation of an Enhanced Human-Robot Interface," in SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme - System Security and Assurance (Cat. No.03CH37483) (Washington, DC, USA: IEEE), 900–905. doi:10.1109/ICSMC.2003.1243929

Jones, E. E., and Gerard, H. B. (1967). eng. *Foundations Of Social Psychology*. New York London Sydney: Wiley. OCLC: 6964475.

Kennedy, J., Baxter, P., Senft, E., and Belpaeme, T. (2016). "Social Robot Tutoring for Child Second Language Learning," in 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (Christchurch, New Zealand: IEEE), 231–238. doi:10.1109/HRI.2016.7451757

Lamont, A. (2011). The Beat Goes on: Music Education, Identity and Lifelong Learning. *Music Educ. Res.* 13, 369–388. doi:10.1080/14613808.2011.638505

Lehmann, A. C., and Ericsson, K. A. (1997). Research on Expert Performance and Deliberate Practice: Implications for the Education of Amateur Musicians and Music Students. *Psychomusicology: A J. Res. Music Cogn.* 16, 40–58. doi:10.1037/h0094068

Lehmann, A. C., Sloboda, J. A., and Woody, R. H. (2007). *Psychology for Musicians: Understanding and Acquiring the Skills*. Oxford; New York: Oxford University Press. OCLC: ocm65197761

MacIntyre, P. D., Schnare, B., and Ross, J. (2018). Self-determination Theory and Motivation for Music. *Psychol. Music* 46, 699–715. doi:10.1177/0305735617721637

Marsh, H. W., Craven, R. G., and Debus, R. (1991). Self-concepts of Young Children 5 to 8 Years of Age: Measurement and Multidimensional Structure. *J. Educ. Psychol.* 83, 377–392. doi:10.1037/0022-0663.83.3.377

McPherson, G. E., and Renwick, J. M. (2001). A Longitudinal Study of Self-Regulation in Children's Musical Practice. *Music Educ. Res.* 3, 169–186. doi:10.1080/14613800120089232

Mellor, D., and Moore, K. A. (2014). The Use of Likert Scales with Children. *J. Pediatr. Psychol.* 39, 369–379. doi:10.1093/jpepsy/jst079

Moore, D. G., Burland, K., and Davidson, J. W. (2003). The Social Context of Musical success: A Developmental Account. *Br. J. Psychol.* 94, 529–549. doi:10.1348/000712603322503088

Nielsen, J., Bærendsen, N. K., and Jessen, C. (2008). *RoboMusicKids–music Education with Robotic Building Blocks*. Banff: IEEE, 149–156.

Papert, S. (1993). eng. *The Children's Machine: Rethinking School in the Age of the Computer*. New York: Basic Books, OCLC: 27264650

Park, H. W., Rosenberg-Kima, R., Rosenberg, M., Gordon, G., and Breazeal, C. (2017). "Growing Growth Mindset with a Social Robot Peer," in Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (Vienna Austria: ACM), 137–145.. doi:10.1145/2909824.3020213

Parncutt, R., and McPherson, G. (2002). English. *The Science & Psychology of Music Performance: Creative Strategies for Teaching and Learning*. Oxford; New York: Oxford University Press. OCLC: 57491519

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educ. Psychol. Rev.* 18, 315–341. doi:10.1007/s10648-006-9029-9

Perugia, G., van Berkel, R., Díaz-Boladeras, M., Català-Mallofré, A., Rauterberg, M., and Barakova, E. (2018). Understanding Engagement in Dementia through Behavior. The Ethographic and Laban-Inspired Coding System of Engagement (ELICSE) and the Evidence-Based Model of Engagement-Related Behavior (EMODEB). *Front. Psychol.* 9, 690. doi:10.3389/fpsyg.2018.00690

Pintrich, P. R., and Schunk, D. H. (2002). Motivation In Education: Theory, Research, and Applications *(Upper Saddle River, N.J: Merrill)*. 2nd ed edn. Englewood Cliffs: Merrill-Prentice Hall.

Rambli, D. R. A., Matcha, W., and Sulaiman, S. (2013). Fun Learning with AR Alphabet Book for Preschool Children. *Proced. Comput. Sci.* 25, 211–219. doi:10.1016/j.procs.2013.11.026

Rassafiani, M. (2009). Is Length of Experience an Appropriate Criterion to Identify Level of Expertise?. *Scand. J. Occup. Ther.* 16, 247–256. doi:10.3109/11038120902795441

Rauscher, F. H., and Hinton, S. C. (2006). The Mozart Effect: Music Listening Is Not Music Instruction. *Educ. Psychol.* 41, 233–238. doi:10.1207/s15326985ep4104_3

Renwick, J. M., and Reeve, J. (2012). *Supporting Motivation in Music Education*. New York: Oxford University Press. doi:10.1093/oxfordhb/9780199730810.013.0009

Riether, N., Hegel, F., Wrede, B., and Horstmann, G. (2012). "Social Facilitation with Social Robots?," in Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12 (Boston, Massachusetts, USA: ACM Press), 41. doi:10.1145/2157689.2157697

Ryan, R. M., and Deci, E. L. (2000). Self-determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *Am. Psychol.* 55, 68–78. doi:10.1037/0003-066X.55.1.68

Sandygulova, A., Johal, W., Zhexenova, Z., Tleubayev, B., Zhanatkyzy, A., Turarova, A., et al. (2020). "CoWriting Kazakh: Learning a New Script with a Robot," in Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. (Cambridge United Kingdom: ACM), 113–120. doi:10.1145/3319502.3374813

Sasfy, J., and Okun, M. (1974). Form of Evaluation and Audience Expertness as Joint Determinants of Audience Effects. *J. Exp. Soc. Psychol.* 10, 461–467. doi:10.1016/0022-1031(74)90014-6

Shiomi, M., Kanda, T., Howley, I., Hayashi, K., and Hagita, N. (2015). Can a Social Robot Stimulate Science Curiosity in Classrooms?. *Int. J. Soc. Robotics* 7, 641–652. doi:10.1007/s12369-015-0303-1

Song, H., Barakova, E., Ham, J., and Markopoulos, P. (2021). *The Impact of Presence of Social Robot and Roles of Robot on Children's Motivation and Performance in Musical Instrument Practicing*. Manuscript in preparation

Song, H., Zhang, Z., Barakova, E. I., Ham, J., and Markopoulos, P. (2020). "Robot Role Design for Implementing Social Facilitation Theory in Musical Instruments Practicing," in Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (Cambridge United Kingdom: ACM), 253–260. doi:10.1145/3319502.3374787

Thresher, J. M. (1964). The Contributions of Carl Orff to Elementary Music Education. *Music Educators J.* 50, 43–48. doi:10.2307/3390084

Tisza, G., and Markopoulos, P. (2021). FunQ: Measuring the Fun Experience of a Learning Activity with Adolescents. *Curr. Psychol.*, 1–21. doi:10.1007/s12144-021-01484-2

Triplett, N. (1898). The Dynamogenic Factors in Pacemaking and Competition. *Am. J. Psychol.* 9, 507. doi:10.2307/1412188

Varela, W., Abrami, P. C., and Upitis, R. (2016). Self-regulation and Music Learning: A Systematic Review. *Psychol. Music* 44, 55–74. doi:10.1177/0305735614554639

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbreton, A. J. A., Freedman-Doan, C., et al. (1997). Change in Children's Competence Beliefs and Subjective Task Values across the Elementary School Years: A 3-year Study. *J. Educ. Psychol.* 89, 451–469. doi:10.1037/0022-0663.89.3.451

Wilgenbusch, T., and Merrell, K. W. (1999). Gender Differences in Self-Concept Among Children and Adolescents: A Meta-Analysis of Multidimensional Studies. *Sch. Psychol. Q.* 14, 101–120. doi:10.1037/h0089000

Williamon, A., and Valentine, E. (2000). Quantity and Quality of Musical Practice as Predictors of Performance Quality. *Br. J. Psychol.* 91, 353–376. doi:10.1348/000712600161871

Woody, R. H. (2004). The Motivations of Exceptional Musicians. *Music Educators J.* 90, 17–21. doi:10.2307/3399950

Xie, L., Antle, A. N., and Motamedi, N. (2008). "Are Tangibles More Fun?: Comparing Children's Enjoyment and Engagement Using Physical, Graphical and Tangible User Interfaces," in Proceedings of the 2nd international conference on Tangible and embedded interaction - TEI '08 (Bonn, Germany: ACM Press), 191. doi:10.1145/1347390.1347433

Zajonc, R. B. (1965). Social Facilitation. *Science* 149, 269–274. doi:10.1126/science.149.3681.269

Zdzinski, S. F. (1996). Parental Involvement, Selected Student Attributes, and Learning Outcomes in Instrumental Music. *J. Res. Music Educ.* 44, 34–48. doi:10.2307/3345412

Zimmerman, B. J. (1989). A Social Cognitive View of Self-Regulated Academic Learning. *J. Educ. Psychol.* 81, 329–339. doi:10.1037/0022-0663.81.3.329

Zimmerman, P. H., Bolhuis, J. E., Willemsen, A., Meyer, E. S., and Noldus, L. P. J. J. (2009). The Observer XT: A Tool for the Integration and Synchronization of Multimodal Signals. *Behav. Res. Methods* 41, 731–735. doi:10.3758/brm.41.3.731

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 in Robotics and AI

ORIGINAL RESEARCH
published: 21 January 2022
doi: 10.3389/frobt.2022.734955

Do Robotic Tutors Compromise the Social-Emotional Development of Children?

Matthijs H. J. Smakman[1,2], Elly A. Konijn[1] and Paul A. Vogt[3]*

[1]Department of Communication Science, VU University Amsterdam, Amsterdam, Netherlands, [2]Institute for Information and Communication Technology, HU University of Applied Sciences Utrecht, Utrecht, Netherlands, [3]School of Communication, Media and IT, Research Group Digital Transformation, Hanze University of Applied Sciences Groningen, Groningen, Netherlands

Social robots are reported to hold great potential for education. However, both scholars and key stakeholders worry about children's social-emotional development being compromised. In aiming to provide new insights into the impact that social robots can have on the social-emotional development of children, the current study interviewed teachers who use social robots in their day-to-day educational practice. The results of our interviews with these experienced teachers indicate that the social robots currently used in education pose little threat to the social-emotional development of children. Children with special needs seem to be more sensitive to social-affective bonding with a robot compared to regular children. This bond seems to have positive effects in enabling them to more easily connect with their human peers and teachers. However, when robots are being introduced more regularly, daily, without the involvement of a human teacher, new issues could arise. For now, given the current state of technology and the way social robots are being applied, other (ethical) issues seem to be more urgent, such as privacy, security and the workload of teachers. Future studies should focus on these issues first, to ensure a safe and effective educational environment for both children and teachers.

Keywords: social robots, child-robot interaction, education, social development, primary school, social skills, bonding, friendship

OPEN ACCESS

Edited by:
Mohamed Chetouani, Université Pierre et Marie Curie, France

Reviewed by:
Konstantinos Tsiakas, Delft University of Technology, Netherlands
Hatice Kose, Istanbul Technical University, Turkey

****Correspondence:***
Matthijs H. J. Smakman
m.h.j.smakman@vu.nl

Specialty section:
This article was submitted to Human-Robot Interaction, a section of the journal Frontiers in Robotics and AI

Received: *01 July 2021*
Accepted: *07 January 2022*
Published: *21 January 2022*

Citation:
Smakman MHJ, Konijn EA and Vogt PA (2022) Do Robotic Tutors Compromise the Social-Emotional Development of Children? Front. Robot. AI 9:734955. doi: 10.3389/frobt.2022.734955

INTRODUCTION

Social robots are gradually being introduced in primary education. They provide new opportunities for improving cognitive outcomes, such as learning a second language (Vogt et al., 2019; Konijn et al., 2021), rehearsing the times tables (Konijn and Hoorn, 2020), learning sign language (Luccio and Gaspari, 2020) and training handwriting (Aktar Mispa and Sojib, 2020). In addition, social robots are used to support motivational and affective elements of learning (e.g., the learner being attentive, receptive, responsive, reflective, or inquisitive) (Belpaeme et al., 2018). Although social robots show potential as learning or teaching companions for children, according to a recent literature review (Johal, 2020), other studies on the use of social robots in education have reported that it is too early to conclude that robots are, for instance, effective as language tutors (van den Berghe et al., 2019), or more effective than human teachers or other types of technology (Woo et al., 2021). Furthermore, both scholars (Sharkey, 2016; Woo et al., 2021) and stakeholders (Smakman et al., 2021a) have voiced concerns related to social robots potentially harming children's social-emotional development.

Social robots differ from other types of robots used in education, such as STEM robots. Other than STEM robots, social robots are designed to take on social roles such as that of a tutor or peer that assists children during their learning process. Having physical embodiment, the option to act (semi-) autonomously, and the capability to interact with humans by following social norms, can be considered as the three defining capacities for social robots (Hegel et al., 2009). Using these capacities, a robot can act as a social entity, such as in the role of a tutor, a peer, or that of a naïve learner (Hood et al., 2015). The feeling that users are socially connected with robots is central to the field of social robotics (Belpaeme et al., 2013).

Children's social-emotional development is not only important during childhood, but also for adulthood and public health, because it is associated with academic performance, substance abuse, mental health, workplace and academic performance (Cherniss, 2000; Denham, 2006; Tremblay, 2020). Children's social-emotional development can be characterized by five domains: 1) social competence, 2) attachment, 3) emotional competence, 4) self-perceived competence, and 5) temperament/ personality (Denham et al., 2009). Milestones in social-emotional development domains differ per developmental period of children. For the purpose of this study, we will focus on the milestones associated with the primary school period. The first domain, social competence, can be defined as a child's ability and effectiveness in social interaction (Rose-Krasnor, 1997). Children's general developmental tasks related to social competence that should be assessed in primary school are the formation of dyadic friendships, solidification of peer status, and general diminution of physical aggression. Related to attachment, children in primary school should begin to balance the connection to parents and peers. The milestones for children in primary school related to emotional competence are the ability to understand complex emotions, such as unique perspective and ambivalence, and to be able to apply cognitive strategies to regulate emotions. Children's self-perception of competence can be defined as "*one's evaluations of one's own abilities, including the child's own assessment of his/her cognitive, physical and social abilities, especially in comparison with those of others*" (Denham et al., 2009, p. 44). During primary school, children's views of their own competence become more complex, earlier notions of self-perceived competence are solidified and social evaluations by peers and teachers become more important (Denham et al., 2009). Lastly, for the domain temperament/ personality, children's personality attributes become increasingly differentiated during primary school. In earlier research, social robots have been reported to potentially influence several aspects of the social-emotional development domains, such as social competence (Peter et al., 2021) and attachment (Coeckelbergh et al., 2016).

Key stakeholders, such as teachers, parents, and policymakers, have also voiced concerns related to the potential social-affective bond that children may develop with a robot (Serholt et al., 2017; Smakman et al., 2020a, Smakman et al., 2020b). They report worries in the field that such a bond could harm children's social-emotional development (Smakman et al., 2021b). Children bonding with robots could lead to children preferring the interaction with robots over that of their human friends and teachers, potentially resulting in the loss of human contact (Sharkey, 2016; Pandey and Gelin, 2017), social isolation (Kennedy et al., 2016), and dehumanization (Serholt et al., 2017). Children could also start to expect too much from robots, which could lead to children ending up feeling deceived or feeling anxious when the robot is absent (Sharkey, 2016). These potential risks related to the social-affective bond that children may develop with a robot might harm the children's social-emotional development. According to a recent study (Pashevich, 2021), it is still unclear what kind of effect social robots might have on the social-emotional development of children.

Children have been reported to perceive social robots as entities with whom they will likely form social relationships (van Straten et al., 2020). What kind of relationships children form with robots is still unclear. For example, children are reported to perceive social robots as potential private tutors (Shin and Kim, 2007), possible rivals (Shin and Kim, 2007), and even friends (Lin et al., 2009). Various scholars argue that this newly perceived bond with technology might influence children's behavior, both positively and negatively. Researchers have found that robots seem able to elicit socially desirable behavior among children, such as sharing, but they may also elicit socially undesirable behavior, such as aggressive behavior (Peter et al., 2021). Children have also been recorded to express bullying behavior towards an educational robot (Kanda et al., 2012) and others have expressed concerns related to the robot becoming a bully or becoming subject to bullying (Diep et al., 2015). What type of children are more susceptible to the influence of the robot on social-emotional domains, however, is still unclear. According to a recent study (Tolksdorf et al., 2021), the influence of individual variables, such as shyness, are still understudied in the field of child-robot interaction (CRI).

Measuring social-emotional development is complex. For each domain of children's social-emotional development, there exist multiple measurement instruments such as the *Rothbart Child Behavior Questionnaire* for emotional competence (Putnam and Rothbart, 2006), and the *Social Skills Rating System* for social competence (Van der Oord et al., 2005). Furthermore, these scales differ per developmental period and pose challenges in their use in longitudinal studies (Denham et al., 2009). Child-robot interaction studies in education are often short-term studies and rarely deploy robots for more than a few days, according to reviews on social robots in classrooms (Rosanda and Starčič, 2019; Woo et al., 2021). Systematic, long-term evaluation of the potential negative impact of social robots' potential on children's social-emotional development is lacking. This might be explained by social robots still being a nascent technology. An accepted approach to evaluate the potential long term (negative) impact of nascent technology is to include stakeholders into the design and evaluation of technology (Friedman et al., 2008).

Teachers are one of the most important stakeholders when implementing social robots in education. They are not only responsible for the learning process in a classroom, but they also play a key role in children's social-emotional development (Denham et al., 2009). They could therefore provide insights into the potential compromising role of social robots. However, in the extant literature on teachers' perspectives on social robots, teachers have had little experience with robots (Van Ewijk et al., 2020; Xia and LeTendre, 2020; Chootongchai et al., 2021). Additionally, researchers have pointed out that the level of experience with robots could influence stakeholders' perspectives (Serholt et al., 2014). People with experience to working with robots are significantly more likely to have a positive attitude towards social robots, compared to people with little to no experience (Smakman et al., 2021a). This makes it hard to evaluate the potential harms and benefits voiced by teachers in earlier studies.

The lack of experience of stakeholders combined with the limited empirical data, make it hard to evaluate the reported potential risks related to children's social-emotional development. Given that studies are often short-term and stakeholders' worries are hard to evaluate, there is a need to examine the impact that social robots have on children's social-emotional development now that social robots are entering day-to-day education for longer periods of time. Therefore, this study aims to assess the impact of social robots in primary education on the social-emotional development of children. To this aim, we conducted in-depth interviews with teachers who have applied social robots in their day-to-day education. These primary school teachers all have a thorough knowledge of the social-emotional development of the children in their classroom, as this is part of their daily job. Therefore, in our opinion, they are most appropriate persons to assess the impact of social robots on children. Besides the impact on children's social-emotional development, we examined which children, according to the teachers, would be more susceptive to social robots, and what the teachers would consider best practices for using social robots responsibly. In the next section, we will first describe our methodology, followed by our results. Thereafter, we will discuss our main findings in light of earlier research and discuss our conclusions.

MATERIALS AND METHODS

Participants

For qualitative research, such as this interview study, participants can best be selected based on their understanding of the phenomenon (Kuper et al., 2008; Creswell and Creswell, 2009). Therefore, via purposeful sampling, participants were selected. The criterion for participants to be included in our study was: being a primary school teacher in the Netherlands with first-hand experience in using social robots in a real-life educational setting. Participants were recruited through newsletters of robotic companies, messages on social media, snowballing (Ghaljaie et al., 2017) and direct e-mails. Nine experienced teachers agreed to participate in our research (Mean age = 36 SD = 10, 8 Female, 1 Male). On average, they had 12 years of working experience, ranging from 1.5 to 35 years. The participants ranked their own experience with robots on a 1–5-point rating scale (1 = having very little experience and 5 = having very much experience). The mean score for the experience with robots was 3.66 (SD = 0.82). In total, the participants supervised/facilitated the child-robot interaction of 2,660 primary school children from all primary school levels/grades. General information about the teachers who participated in the interviews is shown in **Table 1**.

Materials and Measures

In setting up our interview guidelines (Taylor, 2005), we followed the five phases of the framework for the development of a qualitative semi-structured interview guide created by Kallio et al. (2016). First, we established that a semi-structured interview would be a rigorous data collection method in relation to our research question, because it allows the interviewer to improvise follow-up questions based on the teachers' answers and it allows room for participants' verbal expressions. Second, we created an initial set of questions targeting teachers' perspectives on the robot's influence on children's social-emotional development based on existing literature. These questions included four main themes. The first questions were related to the social demographic data of the participant, such as age and gender, because these are shown to influence people's perception of robots (European Commission, Directorate-General for Communication, 2017). The second type of questions was about how the teachers applied the robots in their classroom. These included which robot they used, but also what role the robot was given in the classroom. Earlier research has shown that children react differently to, for example, a robot as a peer, compared to that of a robot as a teacher (Zaga et al., 2015). Furthermore, role switching has also been shown to have potential as a motivational strategy (Ros et al., 2016). The third and fourth themes were related to the possible perceived social-affective bond of children with the robot and its potential influence on children's social-emotional development. After setting up the initial interview protocol, two expert scholars in social robotics reviewed the interview guide to validate the coverage and relevance of the content. Furthermore, as prescribed by Kallio et al. (2016), the feedback of the experts was used to reformulate the questions and to test the implementation. This resulted in the final list of interview questions, which can be found online (https://osf.io/qne96/).

Procedure and Analysis

Over a span of 2 months, from February to April 2021, the data for this study were collected. Due to the COVID19 pandemic, all interviews were conducted online via Microsoft Teams. The interviews started with a short introduction about the purpose of the study, after which the questions started. As mentioned, the interviews were semi-structured (Kallio et al., 2016), which allowed us to deviate somewhat from the formal set of interview questions when needed, and to explore the thoughts and beliefs of participants in more detail. In general, each interview lasted between 45 min and 1 h. At the end of the interview, we inquired whether participants would like to voice any other potentially relevant information related to child development and robots in education. Lastly, we asked participants if they could provide us with names of other teachers who had applied social robots in their education and might be willing to participate in this study. All interviews were recorded,

TABLE 1 | Data on participants in the interviews.

Interview #	Gender	Age	Experience as a teacher (years)	Experience with robots (1–5 scale)	# Children interacting with a robot
1	F	39	14	4	600
2	F	25	3	3	57
3	F	36	13	2	20
4	F	42	10	5	700
5	F	57	35	4	540
6	F	28	7	3	200
7	M	39	12	4	500
8	F	25	1,5	4	25
9	F	35	14	4	18

for which all participants provided active verbal consent. Afterwards, the recordings were transcribed. All transcriptions were then analyzed using an inductive and deductive coding process through a qualitative data analysis application (ATLAS.ti, version 9). To identify patterns within and across the data, we used a thematic analysis method (Braun and Clarke, 2012). First, we coded the text based on the main themes of the interview questions (participant data, use of robots, social-affective bond, and social-emotional development). Thereafter, we randomly read samples of the data and created thematic codes, shown in **Table 1**. We then applied the codes onto new sample texts derived from our interview transcriptions. Using this iterative process, we created our final coding scheme which we applied to all data collected. The themes were coded by a scholar with considerable experience in conducting qualitative studies in social robotics and education. The final coding scheme can be found online (https://osf.io/qne96/). Lastly, the effects of the robots on children derived from the thematic analysis were linked to the appropriate domains of children's social-emotional development reported in the literature (Denham et al., 2009). This was done during a mapping workshop by the first author and two undergraduate students.

RESULTS

All participants had experience with applying humanoid robots in their education, being either with the Nao robot (SoftBank Robotics, 2020) or the Alpha mini-robot (Ubtech, 2021). One participant also had experience with other types of robots, such as the Innobot, Probot, Bluebot, Microbot, and Ozobot. Experience with applying robots ranged from 6 years to a couple of months. The participants had applied the robots in their day-to-day education for teaching children arithmetic, language, geography, presentation skills, physical education, and computational thinking. Eight participants had used the robots as a social entity (as a tutor or peer), sometimes combined with using the robot purely as a tool, such as for learning programming. One participant had used the robot just as a tool for teaching programming. The number of interactions with the robot per child ranged from just one to sixteen times per period of 10 weeks. The time children had spent working with the robot ranged from 15 min to 1 hour per interaction. None of the participants systematically measured the effect of the robot during their lectures. The teachers used robots in all classes of the primary school, which included children from age 4 up to 12 years.

Place in Education

Eight out of nine teachers mentioned that social robots (should) have a place in primary education. They considered the robot a good educational tool, mainly because it can enrich the lessons. *"Some children learn more easily from books, another child learns more easily from a screen with interactivity, and a robot gives an extra dimension to education [. . .] it is one of the means by which you prepare children for a future"* as one teacher indicated. The teachers overall stated that they viewed the robots as additional support for the teacher, or to provide help for solving problems (such as knowledge gaps) by means of targeted help. Teachers had applied the robot in small groups and in one-to-one interaction settings. Most teachers indicated that the robot has a clear novelty effect and that children are fascinated and amazed by the robot. Most of the teachers stated that the children are enthusiastic about the robot and are (more) motivated to work and learn with the robot.

One teacher did not consider social robots to have a place in primary education, for two reasons: 1) because of the high cost and 2) because of a lack of impact in primary education. Although the teacher stated that the robot does create a deeper kind of learning, because of the social interaction, she considered the robot best for special education. In special education, the teaching methods would be more open-minded for using robots and not so restricted and formalized as in regular primary education, according to this teacher. Three other teachers also indicated that the high cost of the NAO robot was an issue. Especially for teaching programming skills, they considered nonsocial or non-humanlike robots cheaper and therefore more appropriate.

Overall, the teachers indicated that the current social robots require a lot of work from the teacher. As one teacher explained: *"It is really labour-intensive for the person who sets up and prepares the robot, and this is still an impeding factor."* Teachers also indicated that it will take some time before other teachers are acquainted with robots because the educational methods change rapidly every few years, which

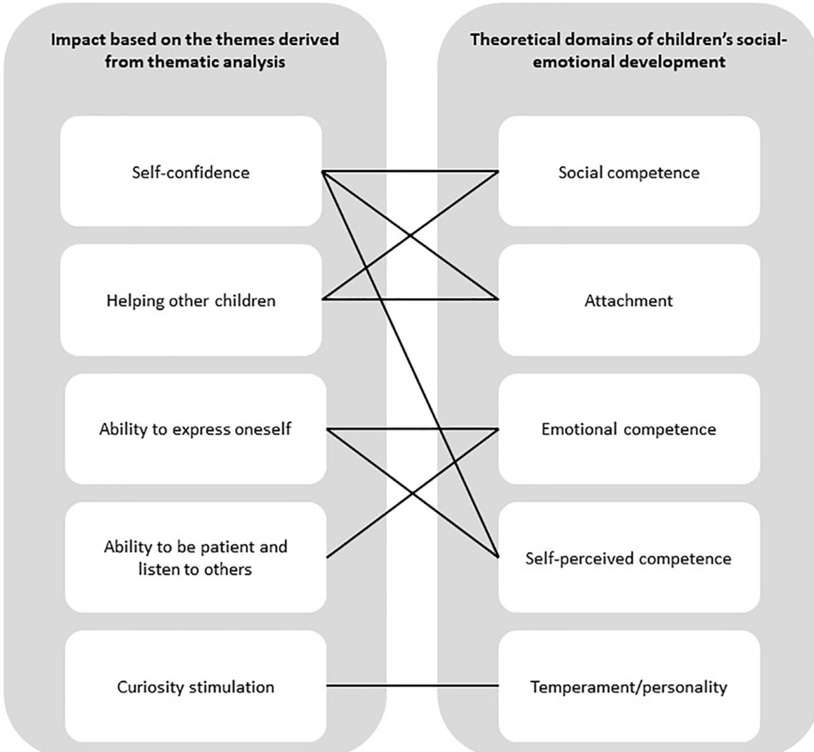


FIGURE 1 | Overview of themes based on the interviews and the linked theoretical constructs of social-emotional development, based on the literature.

also takes time to implement. Furthermore, the lack of evidence that robots are (more) effective makes it hard to convince school management to invest in the implementation of social robots, according to one teacher.

Impact on Social-Emotional Development?

It should first be noted that none of the teachers systematically measured the robot's effect on the children's social-emotional development. Due to the relatively broad age range of the children that interacted with the robot (4–12 years), which covers both the primary school period and the preschool/early childhood period, and because the general developmental tasks that should be assessed in each dimension of social-emotional development differs for each developmental period, we decided to describe the perceived impact based on the themes derived from our thematic analysis (Braun and Clarke, 2012).

All teachers indicated that social robots can have a positive impact on the social-emotional development of children. They reported several examples of how children's social-emotional development could be affected by social robots, such as by boosting children's self-confidence and by increasing children's ability to express themselves. All reported impact was considered positive. Only a few occasions were reported where some (mainly young) children were afraid of the robot. Based on the thematic analysis, we were able to distinguish five positive effects which were reported by the teachers, being: 1) Self-confidence, 2) helping other children, 3) ability to express oneself, 4) ability to be patient and listen to others, and 5) curiosity stimulation. Thereafter, we linked the themes to the appropriate domains of children's social-emotional development, shown in **Figure 1**. In the next sections, we will present the results based on the derived themes and discuss their potential effect on the theoretical domains of children's social-emotional development.

Self-Confidence

Almost half of the teachers reported higher self-confidence as a positive result of child-robot interaction. Children who were shy to talk in public or in groups could give presentations together with the robot, which could bolster the self-confidence of the children. One teacher explained: *"Giving presentations causes a lot of stress in children. I think it is good if you give them a choice, that they can give the presentation, in the first instance, completely by the robot, and then for example together, so that children, perhaps unconsciously, are presenting in front of groups. This way they will get used to it, in a very safe manner [. . .] you actually take away a lot of stress"*. Also, teachers indicated that children who are a bit shy or socially less capable, could become the robot expert of the class, which would boost their self-confidence: *"I could put children who are socially not very strong in the spotlight so that they would become a robot expert. They were then able to teach other children or help the teacher, so they grew in their whole being because of this . . . , this changed their [social] position and place in the group"*, as one teacher explained. Furthermore, teachers reported that children more easily practice subjects they find difficult with the robot because a robot does not judge or laugh at them when they give a wrong answer. This is also reported to create more social interaction between children, as one teacher who used the robots for extra support in language learning described: "*We have a school where several children come from a different culture. They have difficulty speaking Dutch, and they don't speak Dutch at home. They find it difficult to speak in public, and a robot helps them with this and thus helps with their own language development, which also makes it easier for them to make contact with peers. That is what we have seen, it absolutely had an impact*". None of the teachers reported negative outcomes related to the self-confidence of children. Although, some teachers reported practical issues related to the speech of the robot that sometimes lacks the proper pronunciation, especially with longer words.

The capability of the robot to contribute to children's self-confidence can be (in)directly linked to three of the five social-emotional domains. First, the increased social interaction between children, caused by the increased self-confidence of shy children, could lead to the formation of dyadic friendships, which is linked to the social-emotional domain of *social competence*. Furthermore, this could lead to a more balanced connection with their peers, which is related to the social-emotional domain of *attachment*. Lastly, the robot could contribute to the domain of *self-perceived competence*, because it could result in a child's increased ability to assess one's own social abilities in comparison with those of others.

Helping Other Children

Several teachers indicated that they applied the robot to enhance social interaction between children. For example, by giving some children the role of robot expert, they created a new role in the group. According to teachers, this did not only increase children self-confidence (cf. *Self-Confidence*), but it also allowed the robot experts (often the socially weaker children) to more easily interact with other classmates. Also, by letting children work with the robot in small groups, the interaction between children in the groups was stimulated. Furthermore, when the robot was used by multiple groups in sequence, the last group could help the next group when they encountered difficulties. One teacher expressed concerns about when the robot would be used for one-on-one tutoring, which could potentially lower the contact with other children. The teacher considered this as part of the broader trend of (smart)phone use and time spent on a computer, which seems to lower personal, face-to-face contact. However, the teacher could not tell whether the robot caused children to interact less with each other. Likewise, this was not reported by any of the other interviewed teachers.

The option to apply a social robot to stimulate helping behavior can be (in)directly linked to two of the five social-emotional domains. The introduction of social robots, which allows for the creation of new roles in the classroom as indicated by the teachers to stimulate interaction, and can be linked to the domains of *social competence* and *attachment*.

Ability to Express Oneself

Some teachers reported on children who, before the introduction of the robot, would not be willing to talk to the teacher, or did not want to learn. However, after the robot was introduced in the classroom, these children started to talk. First to the robot, and thereafter to the teacher. Teachers said that they expected that children would more easily express certain things to robots than to their teachers. *"I think that a robot could definitely be used for that [emotional support] as well [. . .] because it is something that is a bit further away from you and a bit less personal, so I think it is easier to discuss more difficult things [. . .] and certainly in the social, emotional area,"* as voiced by one of the teachers. Some teachers used the robot as a means to let children talk about their feelings by letting the robot express emotions. This has led to the opportunity to talk about emotional feelings. One teacher compared this to hand puppets that are currently used in the Dutch educational system to start confirmations on difficult subjects, which the teacher considered a similar tool.

Children opening up to a robot about their feelings relates to two of the five social-emotional domains. First, it could allow children to cope with negative emotions, learn about emotions and emotional expressiveness, which is linked to *emotional competence*. Second, it allows for the possibility for children to get more insight into their own social competence, which is related to the domain of *self-perceived competence*.

Ability to be Patient and Listen to Others

Two of the teachers reported on the robot's ability to teach children to be patient and listen more carefully to others. This was mainly caused by the robot's script that did not allow a child to go any faster, according to the teachers. "*You have to keep calm and you also have to keep your impulses in check [. . .] you also have to be careful, children are normally rumbling everywhere, in a manner of speaking, but that is really not possible. So yes, there is really something being asked of them*", as one teacher reported. The teachers indicated that the robot made children listen more to others and wait their turn. However, they also indicated that the robot would need to be in the classroom for longer periods to make a lasting impact on these skills.

The ability to be patient and more carefully listen to others could, in theory, contribute to understanding the unique perspective of others, which can be linked to *emotional competence*.

Curiosity Stimulation

Several teachers indicated that they have seen how robots can stimulate children's curiosity. Most teachers reported on the robot being something "magical" or "special". This made children curious to learn about and from the robot, also for subjects they would otherwise dislike or even avoid. One teacher experienced the following: "*I had one child at that time, who did not want to learn. That does not happen often, but he really did not want to, he had no interest at all in reading or in letters or in math or something else, but that robot that was really it. Once that robot was there, he did everything. That was so special, he did everything he had to do, but not with me, but with the robot. With me, he just* *closed down, but with the robot, he did it all."* The teacher indicated that she did not encounter this behavior often. Other teachers mentioned that they had also experienced how robots stimulated and motivated children, although they voiced that they did not consider the currently limited interactions enough to have a long-lasting effect on children's curiosity.

The ability of the robot to stimulate curiosity can contribute to the social-emotional domain of temperament/personality. By stimulating children's curiosity, they could become more encouraged to follow and experience aspects that suit their personality, which can be linked to the *temperament/personality* domain.

In summary, the teachers expressed five ways by which social robots can impact the theoretical domains of children's social-emotional development, which is illustrated in **Figure 1**. Children potentially getting attached to the robot was a topic that came up regularly during the interviews. Therefore, we decided to discuss attachment as a topic separately in the next section.

Attachment

Almost half of the teachers indicated that children can feel emotionally attached to a robot. Some indicated that this attachment would not be different from how children attach to other objects children like, such as video games and toys. One teacher saw a child with bonding problems getting emotionally attached to the robot, but did not encounter this with other (typically developing) children. Some teachers indicated that while young children could feel attached to the robot, older children, around the age of 11-12, would consider the robots merely as a tool.

Another teacher reported on a child talking about the robot as his best friend, while other participants indicated that they have seen children interact with the robot as if it were their buddy. In several interviews, teachers indicated that children showed a kind of empathy and affection towards the robot. As one teacher experienced: "*They [children] also immediately asked when he [the robot] would come back, and everyone wanted to take care of it, you really noticed that the care aspect really came up there. Such that it had actually become a kind of a buddy."* Another teacher indicated to be concerned that children would view the robot as a best friend, however, this teacher did not encounter this in her own classroom. Furthermore, several teachers indicated that for children to become attached to a robot, the robot would have to be present much more often than is possible in the current educational system.

What Would be Considered "Too Attached"?

When asked for signals that would indicate that children are too attached to the robot, teachers expressed two main indicators: 1) when it results in less contact with their human peers, and 2) when children would get upset when the robot was not around. However, four teachers indicated explicitly that they have not encountered this in their classes, and that the way robots are being applied nowadays poses little risk for children to become too attached. *"In the current education you don't get it [attachment issues] very quickly, only if you always have a robot in class"* and *"I see few risks in the way in which we now use robots"* as explained by two other teachers. The other five teachers did mention encountering attachment issues in their classes.

Although the teachers did not encounter children becoming too attached to the social robots, this might be due to the short interaction time and the limited number of interactions children had with the robot. Therefore, we continued to further ask the teachers on what type of children would be more susceptible to getting attached to social robots.

Children who are More Susceptible to Getting Attached to Social Robots

The current literature does not provide a solid basis for deriving insights into what kind of children would be more susceptive to getting attached to social robots. To gain more insight into which children might be at risk to become 'too attached' to a social robot, we conducted a thematic analysis to differentiate between types of children based on the interview transcripts (Braun and Clarke, 2012). The teachers expressed four types of children who would be more susceptible to getting attached to social robots.

- The first type, indicated by seven of the nine teachers, is timid, socially less strong, and could have an autism spectrum disorder (ASD). However, regarding ASD, it should be noted that one teacher explicitly stated that these usually are children of which the teachers *think* they have ASD because it is mostly not yet diagnosed at this young age. Indeed, a number of studies reported successful interactions of social robots specifically focusing on children with ASD (e.g., Huijnen et al., 2016; Di Nuovo et al., 2020).
- The second type of children concerns children who are interested in science and engineering. "*The children who are just very interested in robots and programming*", as one teacher explained. This is in line with common applications of robots for STEM education (e.g., Ahmad et al., 2020).
- The third type of children that can be considered more sensitive for the robot's interaction, as indicated by two teachers, are children who are underachievers on a certain subject, such as language learning or math. Studies indeed reported good results for language learning (Vogt et al., 2019; Konijn et al., 2021) or rehearsing the times tables (Konijn and Hoorn, 2020).
- The fourth and final type of children who are more sensitive to social robots are children with special needs, such as children with attention deficit hyperactivity disorder (ADHD), highly sensitive children, highly gifted children, and children sensitive to game addiction. Seven teachers indicated that these children can be considered more sensitive for child-robot interaction in education: "*The children who have a certain need [...] children with ADHD, or just children who are highly gifted, they could be attracted in a certain way if it suits them, and then there are many possibilities to work with this,"* as explained by one teacher. In earlier studies the potential for children with ADHD have been discussed before (e.g., Fridin and Yaakobi, 2011)

TABLE 2 | Best practices and success factors for applying social robots in primary education.

#	Title	Description
1	Apply when needed	Make sure there is a clear *why* for applying social robots, robots are means not ends. Social robots are considered to be an addition to the teacher, not a replacement. When applying the robot every day, the novelty effect can wear off. Use social robots for a specific aim or goal
2	Teacher stays involved	The role of the teachers stays very important, he/she should be present during the child-robot interaction, or at least close by. Also, the teacher can judge which children potentially get too attached to the robot, and which children would benefit most from the interaction. This might lead to an increase in the number of teaching assistants needed to facilitate the robot interaction
3	Proper introduction	Teachers should pay specific attention to the introduction of the robot. Children should first be told what a robot is, and what is it going to do, before they start to interact with a robot
4	Small groups	Learning with robots is best done in small groups. This not only allows children to continue communicating with their peers, but it can also stimulate children to interact with each other and not get socially isolated
5	Vertical groups	Let children of different age groups work together with the robot, make use of the older, more experienced children to introduce and guide younger children
6	Separate room	When a small group of children is working with the robot, this is distracting for the other children in the classroom. Therefore, the robot should not be in the same room as where other children are who do not work with the robot
7	Team effort and mindset	For robots to be sustainably implemented in schools, the technology needs to have the support of the teacher-team including the school management. A teacher in the role of a robot ambassador can be appointed to introduce the robot to other teachers, making it easier to implement the robot
8	Parents	The parents of the children should be informed pro-actively by the schools when social robots are going to be used. This is the responsibility of the school

The teachers expressed several best practices to ensure that these types of children would not get too attached to the robot. The best practices expressed also included general remarks on how social robots could be implemented in a responsible way, according to these experienced teachers. In the next section, we present these findings.

Best Practices and Success Factors for Child-Robot Interaction in Education

The interviewed teachers reported about what they considered best practices and success factors when applying social robots in primary education. In total, they reported eight best practices and success factors for applying social robots in primary education. To provide an overview of these best practices and their description we present them in **Table 2**.

DISCUSSION

The main goal of this interview study was to examine whether social robots in primary education compromise the social-emotional development of children. Therefore, we interviewed primary school teachers who supervised the child-robot interaction of more than 2,600 unique children in a real-life school environment. Nearly all child-robot interactions reported by our interviewees were one-on-one or small group interactions in which a humanoid robot took the role of a tutor or peer. Each robot was used for teaching children a specific subject or skill in a school environment.

The main finding of our study is that the participating teachers experienced no negative effects on the social-emotional development of children caused by the child-robot interactions that would have a lasting negative impact. In contrast, teachers expressed seeing five positive effects of social robots related to the social-emotional development of their pupils, being 1) increased self-confidence, 2) helping other children, 3) increased ability to express oneself, 4) increased ability to be patient and listen to others, and 5) curiosity stimulation. These five themes could be linked to all domains of children's social development reported in developmental literature, as discussed in the introduction and summarized in **Figure 1**.

The social robots seemed especially useful for introducing the learning by teaching paradigm (Fiorella and Mayer, 2013). This allows for some children to take on new roles, such as that of an expert. This can have a positive effect on children's social-emotional development. For example, by giving children an expert role, or by letting experienced groups help other groups. Novel technologies, such as social robots, seem appropriate to support children in such roles. The robot's impact on the children's ability to be patient and to listen carefully was reported to be caused mainly by the current state of technology that does not allow children to respond quickly, and due to the intonation of the robot which is sometimes lacking. Given that automatic speech recognition based on child-robot interaction has been shown to be a complex issue (Kennedy et al., 2017), it is unlikely that robots will be able to respond quickly to children's verbal reactions in the near future. Therefore, we consider that the robot's positive impact on children's ability to be patient and to listen will remain for the foreseeable future. However, teachers indicated that they wondered whether the effect on children's ability to be patient and to listen would impact the children in the long run. The other three effects, increased self-confidence, ability to express oneself, and curiosity stimulation, seem all specifically useful for children with special needs.

Four types of children were identified by the interviewed teachers, three of whom could specifically benefit from social robots and be receptive to interacting with a social robot. These children are considered to have special needs, either the timid, socially less strong children potentially with ASD, underachievers, or children with other special needs, such as ADHD or attachment issues. According to the teachers, these children could potentially benefit the most from social robots in education when it comes to their social-emotional development and are indeed often addressed in studies (e.g., Fridin and Yaakobi, 2011; Huijnen et al., 2016; Konijn and Hoorn, 2020). As a downside, the interviewed teachers reported that these children might get more attached to the robot in the long run, which could, in theory, lead to less human contact and children getting upset when the robot would not be around. However, this has not been observed by our teachers, and they further indicated that the robot would need to be present much more for this to occur.

To ensure that some children will not get too attached to the robot, teachers have indicated that they should supervise the child-robot interaction, or at least be close by. The teachers in our study mentioned that applying social robots in education is labor-intensive, and requires time and effort to use and implement. This is in line with another study reporting about teachers being worried that social robots would increase the workload of teachers (Reich-Stiebert and Eyssel, 2016). A recent review on robots in classrooms came to similar results, concluding that "the current generation of commercially available robots, like NAO or Pepper, do not have sufficient programming to be readily integrated into classrooms without extensive support and resource mobilization" (Woo et al., 2021, p.9).

The comparison of the bond between children and robots to the bond between children and other humans might not be the best way forward. Although some children seem to behave as if they are friends with a robot (Fior et al., 2010), robots are still a different entity. When comparing human-robot interaction to interaction between humans, Black (2019) argues against developing empathy with robots because children cannot experience the kind of affect toward robots that they develop with other humans, such as their human peers and teachers. However, if we use the robot to simulate human interaction, by letting children work together, this doesn't seem to be a big problem. Furthermore, for social robots to be able to support children in primary education, there seems to be no need for very humanlike robots with extensive empathy capabilities; current studies on the use of social robots in education do, most of the time, not use very humanlike robots with extensive empathy capabilities, and still show promising results (e.g., Konijn and Hoorn, 2020). One might argue that robots need extensive empathy capabilities for teaching social skills to children who cannot learn these with their human peers because of disorders, such as ASD. Although humanoid robots with extensive empathy capabilities might help this specific group of children, there seems little reason to equip robots with far-reaching human embodiment when it comes to assisting regular children in their school process.

The social bond between child and robot challenges the fundamentals of friendship and relationships, according to Richards and Calvert (2017). However, according to the teachers in our study, such social bonds are infrequent and similar to the bond children have with other technologies or artefacts, such as smartphones and (hand) puppets. Thus, the negative impact of social robots on the fundamentals of friendship and relationships, for now, seems limited.

Other researchers have found that robots seem able to elicit socially desirable behavior among children, such as sharing (Peter et al., 2021). However, according to the same researchers, this may also apply to socially undesirable behavior, such as aggressive behavior (Peter et al., 2021). Children have been recorded to express bullying behavior towards an educational robot (Kanda et al., 2012). Others have also expressed concerns related to the robot becoming a bully or becoming subject to bullying (Diep et al., 2015). However, following the best practices of the participants in our study, when teachers stay involved in the child-robot interaction, this scenario seems unlikely. Teachers or teaching assistants could intervene when such undesirable behavior occurs. Nevertheless, the results of other researchers emphasize the importance to be careful in how robots are presented to children because robots (in videos) have been shown to negatively influence children's pro-social behavior and willingness to share resources in an experimental setting (Nijssen et al., 2021).

The participating teachers did not report major privacy issues related to the child-robot interaction, except one related to IT security, and they did not use extensive personalized data collection by the robot. This might be due to the relatively simple, not highly personalized child robot interaction currently used in schools. In other studies, privacy has been reported to be a major issue related to social robots in education (Sharkey, 2016; Smakman et al., 2021a). Data collection allows personalized interaction, which is one of the key benefits, according to scholars (Kanda et al., 2012; Shimada et al., 2012; Jones et al., 2017; Jones and Castellano, 2018; Woo et al., 2021). Although the teachers in our study did not report on major privacy issues, given the need for data collection for personalized learning, we consider the issue crucial for integrating social robots in education in a responsible way and should therefore be subject for further research.

One limitation of this study is that, although the participants had experience with using a social robot in their day-to-day education and supervised the child-robot interaction of over 2,600 unique children, the total number of participants was limited. However, given that all participants had experience with using a social robot in their day-to-day education, combined with the large number of unique children they supervised, they still provide valuable insights into the currently observed effects of social robots on children. The gender distribution was unequally balanced, with only one male participating teacher. However, this can be considered a reflection of the gender distribution in Dutch primary education, where approximately 80% is female (Traag, 2018). It should also be noted that this study was carried out solely in the Netherlands, therefore the results may differ in other countries. Furthermore, none of the teachers systematically measured the robot's effect on the social-emotional development of children. The evaluations in this study are solely based on the teachers' previous experiences and observations. The experiences of these teachers could differ from how children experienced the robot interaction. Further studies could compare the perceptions of children to the perceptions of their teachers. Future studies in child-robot interaction could also include the Social Skills Rating System (SSRS) or the social Skills Improvement System-Rating Scales (SSIS-RS) (Gresham et al., 2011), to systematically measure the impact of social robots in children's development.

In conclusion, our study indicates that the social robots currently used in education pose little threat to the social-emotional development of children according to teachers who applied these robots in their day-to-day education. Children with special needs seem to be more sensitive to social bonding with a robot compared to regular children. However, this social-affective bond seems to have more positive effects enabling them to more easily connect with their human peers and teachers.

Given that the best practices reported in this study are taken into account, we consider that social robots pose more benefits than harms concerning the social-emotional development of children. However, when robots are being introduced more regularly, daily, without the involvement of a human teacher, new issues could arise. For now, given the current state of technology and the way social robots are being applied, other (ethical) issues seem to be more urgent, such as privacy and security issues, and the workload of teachers.

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 in the article/**Supplementary Material**.

ETHICS STATEMENT

The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Conceptualization, MS and PV. EK; data curation, MS; formal analysis, MS; funding acquisition, MS. EK; investigation, MS; methodology, MS, EK, and PV; project administration, MS resources, MS and EK; software, MS; supervision, EK and PV; validation, EK and PV; visualization, MS; writing original draft, MS; writing review and editing, EK, PV, and MS.

FUNDING

Research supported by The Dutch Research Council (NWO), project number 023.010.066.

ACKNOWLEDGMENTS

We thank all participants in our study. We also would like to thank the robotic companies that included our request to participate in our study in their newsletter.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frobt.2022.734955/full#supplementary-material

REFERENCES

Ahmad, M. I., Khordi-moodi, M., and Lohan, K. S. (2020). "Social Robot for STEM Education," in Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction HRI '20 (New York, NY, USA: Association for Computing Machinery), 90–92. doi:10.1145/3371382.3378291

Aktar Mispa, T., and Sojib, N. (2020). "Educational Robot Kiddo Learns to Draw to Enhance Interactive Handwriting Scenario for Primary School Children," in 2020 3rd International Conference on Intelligent Robotic and Control Engineering (IRCE), 87–91. doi:10.1109/IRCE50905.2020.9199254

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., et al. (2013). Multimodal Child-Robot Interaction: Building Social Bonds. *J. Hum.-Robot Interact.* 1, 33–53. doi:10.5898/jhri.1.2.belpaeme

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social Robots for Education: A Review. *Sci. Robot.* 3, eaat5954. doi:10.1126/scirobotics.aat5954

Black, D. (2019). Machines with Faces: Robot Bodies and the Problem of Cruelty. *Body Soc.* 25, 3–27. doi:10.1177/1357034X19839122

Braun, V., and Clarke, V. (2012). "Thematic Analysis," in APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological APA handbooks in psychology® (Washington, DC, US: American Psychological Association), 57–71. doi:10.1037/13620-004

Cherniss, C. (2000). *"Social and Emotional Competence in the Workplace," in* The handbook of Emotional Intelligence: Theory, Development, Assessment, and Application at home, School, and in the Workplace. San Francisco, CA, US: Jossey-Bass, 433–458.

Chootongchai, S., Songkram, N., and Piromsopa, K. (2021). Dimensions of Robotic Education Quality: Teachers' Perspectives as Teaching Assistants in Thai Elementary Schools. *Educ. Inf. Technol.* 26, 1387–1407. doi:10.1007/s10639-019-10041-1

Coeckelbergh, M., Pop, C., Simut, R., Peca, A., Pintea, S., David, D., et al. (2016). A Survey of Expectations About the Role of Robots in Robot-Assisted Therapy for Children with ASD: Ethical Acceptability, Trust, Sociability, Appearance, and Attachment. *Sci. Eng. Ethics.* 22, 47–65. doi:10.1007/s11948-015-9649-x

Creswell, J. W., and Creswell, J. D. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 3rd ed. Thousand Oaks, CA: Sage.

Denham, S. A. (2006). Social-Emotional Competence as Support for School Readiness: What Is it and How Do We Assess it? *Early Education Development.* 17, 57–89. doi:10.1207/s15566935eed1701_4

Denham, S. A., Wyatt, T. M., Bassett, H. H., Echeverria, D., and Knox, S. S. (2009). Assessing Social-Emotional Development in Children from a Longitudinal Perspective. *J. Epidemiol. Community Health.* 63, i37–i52. doi:10.1136/jech.2007.070797

Di Nuovo, A., Bamforth, J., Conti, D., Sage, K., Ibbotson, R., Clegg, J., et al. (2020). "An Explorative Study on Robotics for Supporting Children with Autism Spectrum Disorder during Clinical Procedures," in Companion of the 2020

ACM/IEEE International Conference on Human-Robot Interaction HRI ’20 (New York, NY, USA: Association for Computing Machinery), 189–191. doi:10.1145/3371382.3378277

Diep, L., Cabibihan, J.-J., and Wolbring, G. (2015). “Social Robots,” in Proceedings of the 3rd 2015 Workshop on ICTs for improving Patients Rehabilitation Research Techniques (Lisbon, Portugal: ACM Press), 160–163. doi:10.1145/2838944.2838983

European Commission, Directorate-General for Communication (2017). Special Eurobarometer 460: Attitudes towards the Impact of Digitisation and Automation on Daily Life. Available at: https://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_382_en.pdf (Accessed December 2, 2020).

Fior, M., Nugent, S., Beran, T. N., Ramírez-Serrano, A., and Kuzyk, R. (2010). Children’s Relationships with Robots: Robot Is Child’s New Friend. *J. Phys. Agents.* 4, 9–17. doi:10.14198/JoPha.2010.4.3.02

Fiorella, L., and Mayer, R. E. (2013). The Relative Benefits of Learning by Teaching and Teaching Expectancy. *Contemp. Educ. Psychol.* 38, 281–288. doi:10.1016/j.cedpsych.2013.06.001

Fridin, M., and Yaakobi, Y. (2011). “Educational Robot for Children with ADHD/ADD, Architectural Design,” in Proceedings of the International Conference on Computational Vision and Robotics (Bhubaneswar, India.

Friedman, B., Kahn, P. H., and Borning, A. (2008). “Value Sensitive Design and Information Systems,” in *The Handbook of Information and Computer Ethics*. Editors K. E. Himma and H. T. Tavani (Hoboken, New Jersey: John Wiley & Sons), 69–101.

Ghaljaie, F., Naderifar, M., and Goli, H. (2017). Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research. *Strides Development Med. Education.* 14. doi:10.5812/SDME.67670

Gresham, F. M., Elliott, S. N., Vance, M. J., and Cook, C. R. (2011). Comparability of the Social Skills Rating System to the Social Skills Improvement System: Content and Psychometric Comparisons Across Elementary and Secondary Age Levels. *Sch. Psychol. Q.* 26, 27–44. doi:10.1037/a0022662

Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., and Sagerer, G. (2009). “Understanding Social Robots,” in 2009 Second International Conferences on Advances in Computer-Human Interactions, 169–174. doi:10.1109/ACHI.2009.51

Hood, D., Lemaignan, S., and Dillenbourg, P. (2015). “When Children Teach a Robot to Write,” in Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction HRI ’15 (Portland, Oregon, USA: ACM), 83–90. doi:10.1145/2696454.2696479

Huijnen, C. A. G. J., Lexis, M. A. S., and de Witte, L. P. (2016). Matching Robot KASPAR to Autism Spectrum Disorder (ASD) Therapy and Educational Goals. *Int. J. Soc. Robotics.* 8, 445–455. doi:10.1007/s12369-016-0369-4

Johal, W. (2020). Research Trends in Social Robots for Learning. *Curr. Robot Rep.* 1, 75–83. doi:10.1007/s43154-020-00008-3

Jones, A., Bull, S., and Castellano, G. (2017). "I Know that Now, I’m Going to Learn This Next" Promoting Self-Regulated Learning with a Robotic Tutor. *Int. J. Soc. Robotics.* 10, 439–454. doi:10.1007/s12369-017-0430-y

Jones, A., and Castellano, G. (2018). Adaptive Robotic Tutors that Support Self-Regulated Learning: A Longer-Term Investigation with Primary School Children. *Int. J. Soc. Robotics.* 10, 357–370. doi:10.1007/s12369-017-0458-z

Kallio, H., Pietilä, A.-M., Johnson, M., and Kangasniemi, M. (2016). Systematic Methodological Review: Developing a Framework for a Qualitative Semi-structured Interview Guide. *J. Adv. Nurs.* 72, 2954–2965. doi:10.1111/jan.13031

Kanda, T., Shimada, M., and Koizumi, S. (2012). “Children Learning with a Social Robot,” in Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI ’12 (Boston, Massachusetts, USA: ACM Press), 351. doi:10.1145/2157689.2157809

Kennedy, J., Lemaignan, S., and Belpaeme, T. (2016). “The Cautious Attitude of Teachers towards Social Robots in Schools,” in Robots 4 Learning Workshop at IEEE RO-MAN 2016 (New York, NY, USA: IEEE).

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., et al. (2017). “Child Speech Recognition in Human-Robot Interaction,” in Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction HRI ’17 (Vienna, Austria: Association for Computing Machinery), 82–90. doi:10.1145/2909824.3020229

Konijn, E. A., and Hoorn, J. F. (2020). Robot Tutor and Pupils’ Educational Ability: Teaching the Times Tables. *Comput. Education.* 157, 103970. doi:10.1016/j.compedu.2020.103970

Konijn, E. A., Jansen, B., Mondaca Bustos, V., Hobbelink, V. L. N. F., and Preciado Vanegas, D. (2021). Social Robots for (Second) Language Learning in (Migrant) Primary School Children. *Int. J. Soc. Robotics.* doi:10.1007/s12369-021-00824-3

Kuper, A., Lingard, L., and Levinson, W. (2008). Critically Appraising Qualitative Research. *BMJ.* 337, a1035. doi:10.1136/bmj.a1035

Lin, Y.-C., Liu, T.-C., Chang, M., and Yeh, S.-P. (2009). “Exploring Children’s Perceptions of the Robots,” in Learning by Playing. Game-based Education System Design and Development Lecture Notes in Computer Science. Editors R. Kuo, C. Kinshuk, G. D. Chen, and M. Hirose (Springer Berlin Heidelberg), 512–517. doi:10.1007/978-3-642-03364-3_63

Luccio, F. L., and Gaspari, D. (2020). “Learning Sign Language from a Sanbot Robot,” in Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good GoodTechs ’20 (New York, NY, USA: Association for Computing Machinery), 138–143. doi:10.1145/3411170.3411252

Nijssen, S. R. R., Müller, B. C. N., Bosse, T., and Paulus, M. (2021). You, Robot? the Role of Anthropomorphic Emotion Attributions in Children’s Sharing with a Robot. *Int. J. Child-Computer Interaction.* 30, 100319. doi:10.1016/j.ijcci.2021.100319

Pandey, A. K., and Gelin, R. (2017). “Humanoid Robots in Education: A~Short~Review,” in *Humanoid Robotics: A Reference*. Editors A. Goswami and P. Vadakkepat (Dordrecht: Springer Netherlands), 1–16. doi:10.1007/978-94-007-7194-9_113-1

Pashevich, E. (2021). Can Communication with Social Robots Influence How Children Develop Empathy? Best-Evidence Synthesis. *AI Soc.* doi:10.1007/s00146-021-01214-z

Peter, J., Kühne, R., and Barco, A. (2021). Can Social Robots Affect Children’s Prosocial Behavior? An Experimental Study on Prosocial Robot Models. *Comput. Hum. Behav.* 120, 106712. doi:10.1016/j.chb.2021.106712

Putnam, S. P., and Rothbart, M. K. (2006). Development of Short and Very Short Forms of the Children’s Behavior Questionnaire. *J. Personal. Assess.* 87, 102–112. doi:10.1207/s15327752jpa8701_09

Reich-Stiebert, N., and Eyssel, F. (2016). “Robots in the Classroom: What Teachers Think About Teaching and Learning with Education Robots,” in *Social Robotics Lecture Notes in Computer Science* (Cham: Springer), 671–680. doi:10.1007/978-3-319-47437-3_66

Richards, M. N., and Calvert, S. L. (2017). “Media Characters, Parasocial Relationships, and the Social Aspects of Children’s Learning Across Media Platforms,” in *Media Exposure during Infancy and Early Childhood*. Editors R. Barr and D. N. Linebarger (Cham: Springer International Publishing), 141–163. doi:10.1007/978-3-319-45102-2_9

Ros, R., Oleari, E., Pozzi, C., Sacchitelli, F., Baranzini, D., Bagherzadhalimi, A., et al. (2016). A Motivational Approach to Support Healthy Habits in Long-Term Child-Robot Interaction. *Int. J. Soc. Robotics.* 8, 599–617. doi:10.1007/s12369-016-0356-9

Rosanda, V., and Starčič, A. I. (2019). A Review of Social Robots in Classrooms:. *Education and Self Development.* 14, 93–106. doi:10.26907/esd14.3.09

Rose-Krasnor, L. (1997). The Nature of Social Competence: A Theoretical Review. *Soc. Development.* 6, 111–135. doi:10.1111/j.1467-9507.1997.tb00097.x10.1111/1467-9507.00029

Serholt, S., Barendregt, W., Leite, I., Hastie, H., Jones, A., Paiva, A., et al. (2014). Teachers’ Views on the Use of Empathic Robotic Tutors in the Classroom. *IEEE.*, 955–960. doi:10.1109/ROMAN.2014.6926376

Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S., et al. (2017). The Case of Classroom Robots: Teachers’ Deliberations on the Ethical Tensions. *AI Soc.* 32, 613–631. doi:10.1007/s00146-016-0667-2

Sharkey, A. J. C. (2016). Should We welcome Robot Teachers? *Ethics Inf. Technol.* 18, 283–297. doi:10.1007/s10676-016-9387-z

Shimada, M., Kanda, T., and Koizumi, S. (2012). “How Can a Social Robot Facilitate Children’s Collaboration?” *Social Robotics*. Editors S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, and M.-A. Williams (Berlin, Heidelberg: Springer Berlin Heidelberg), 98–107. doi:10.1007/978-3-642-34103-8_10

Shin, N., and Kim, S. (2007). “Learning about, from, and with Robots: Students’ Perspectives,” in 16th IEEE International Conference on Robot & Human Interactive Communication (Jeju, South Korea: IEEE), 1040–1045. doi:10.1109/ROMAN.2007.4415235

Smakman, M., Berket, J., and Konijn, E. A. (2020a). “The Impact of Social Robots in Education: Moral Considerations of Dutch Educational Policymakers,” in 2020

29th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) (Napels, Italy: IEEE). doi:10.1109/ro-man47096.2020.9223582

Smakman, M., Jansen, B., Leunen, J., and Konijn, E. A. (2020b). “Acceptable Social Robots in Education: A Value Sensitive Parent Perspective,” in Proceedings of INTED2020 Conference (Valencia, Spain, 7946–7953. doi:10.21125/inted.202010.21125/inted.2020.2161

Smakman, M., Konijn, E. A., Vogt, P., and Pankowska, P. (2021a). Attitudes towards Social Robots in Education: Enthusiast, Practical, Troubled, Sceptic, and Mindfully Positive. *Robotics.* 10, 24. doi:10.3390/robotics10010024

Smakman, M., Vogt, P., and Konijn, E. A. (2021b). Moral Considerations on Social Robots in Education: A Multi-Stakeholder Perspective. *Comput. Education.* 174, 104317. doi:10.1016/j.compedu.2021.104317

SoftBank Robotics (2020). NAO the Humanoid and Programmable Robot | SoftBank Robotics. Available at: https://www.softbankrobotics.com/emea/en/nao (Accessed September 16, 2020).

Taylor, M. C. (2005). “Interviewing,” in *Qualitative Research in Health Care*. Editor I. Holloway (Maidenhead, England: McGraw-Hill Education), 39–55.

Tolksdorf, N. F., Viertel, F. E., and Rohlfing, K. J. (2021). Do Shy Preschoolers Interact Differently When Learning Language With a Social Robot? An Analysis of Interactional Behavior and Word Learning. *Front. Robot. AI.* 8, 676123. doi:10.3389/frobt.2021.676123

Traag, T. (2018). *Leerkrachten in Het Basisonderwijs*. Den Haag, Netherlands: CBS.

Tremblay, R. E. (2000). The Development of Aggressive Behaviour during Childhood: What Have We Learned in the Past century? *Int. J. Behav. Development.* 24, 129–141. doi:10.1080/016502500383232

Ubtech (2021). Alpha Mini. *UBTECH Robotics.* Available at: https://www.ubtrobot.com/products/alpha-mini (Accessed May 3, 2021).

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., Van der Ven, S., and Leseman, P. (2019). Social Robots for Language Learning: A Review. *Rev. Educ. Res.* 89, 259–295. doi:10.3102/0034654318821286

Van der Oord, S., Van der Meulen, E. M., Prins, P. J. M., Oosterlaan, J., Buitelaar, J. K., and Emmelkamp, P. M. G. (2005). A Psychometric Evaluation of the Social Skills Rating System in Children with Attention Deficit Hyperactivity Disorder. *Behav. Res. Ther.* 43, 733–746. doi:10.1016/j.brat.2004.06.004

Van Ewijk, G., Smakman, M., and Konijn, E. A. (2020). “Teachers’ Perspectives on Social Robots in Education,” in Proceedings of the Interaction Design and Children Conference (London, UK: ACM), 273–280. doi:10.1145/3392063.3394397

van Straten, C. L., Peter, J., and Kühne, R. (2020). Child-Robot Relationship Formation: A Narrative Review of Empirical Research. *Int. J. Soc. Robotics.* 12, 325–344. doi:10.1007/s12369-019-00569-0

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., et al. (2019). “Second Language Tutoring Using Social Robots: A Large-Scale Study,” in 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 497–505. doi:10.1109/HRI.2019.8673077

Woo, H., LeTendre, G. K., Pham-Shouse, T., and Xiong, Y. (2021). The Use of Social Robots in Classrooms: A Review of Field-Based Studies. *Educ. Res. Rev.* 33, 100388. doi:10.1016/j.edurev.2021.100388

Xia, Y., and LeTendre, G. (2020). Robots for Future Classrooms: A Cross-Cultural Validation Study of "Negative Attitudes toward Robots Scale" in the U.S. Context. *Int. J. Soc. Robotics.* 13, 703–714. doi:10.1007/s12369-020-00669-2

Zaga, C., Lohse, M., Truong, K. P., and Evers, V. (2015). “The Effect of a Robot’s Social Character on Children’s Task Engagement: Peer versus Tutor,” in *Social Robotics*. Editors A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi (Cham: Springer International Publishing), 704–713. doi:10.1007/978-3-319-25554-5_70

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 Robotics and AI

TYPE Original Research
PUBLISHED 25 August 2022
DOI 10.3389/frobt.2022.666736

Check for updates

OPEN ACCESS

EDITED BY
Paul Vogt,
Hanze University of Applied Sciences,
Netherlands

REVIEWED BY
Yeshasvi Tirupachuri,
Italian Institute of Technology (IIT), Italy
Emmanuel Senft,
University of Wisconsin-Madison,
United States

*CORRESPONDENCE
Wafa Johal,
wafa.johal@unimelb.edu.au

SPECIALTY SECTION
This article was submitted to Human-Robot Interaction,
a section of the journal
Frontiers in Robotics and AI

RECEIVED 10 February 2021
ACCEPTED 30 June 2022
PUBLISHED 25 August 2022

CITATION
Johal W, Gatos D, Yantac AE and Obaid M (2022), Envisioning social drones in education.
Front. Robot. AI 9:666736.
doi: 10.3389/frobt.2022.666736

Envisioning social drones in education

Wafa Johal[1]*, Doğa Gatos[2], Asim Evren Yantac[2] and Mohammad Obaid[3]

[1]School of Computer Science and IS, Faculty of Engineering and IT, University of Melbourne, Parkville, VIC, Australia, [2]Media and Visual Arts Department, Koç University, Istanbul, Turkey, [3]Interaction Design Unit, Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

Education is one of the major application fields in social Human-Robot Interaction. Several forms of social robots have been explored to engage and assist students in the classroom environment, from full-bodied humanoid robots to tabletop robot companions, but flying robots have been left unexplored in this context. In this paper, we present seven online remote workshops conducted with 20 participants to investigate the application area of Education in the Human-Drone Interaction domain; particularly focusing on what roles a social drone could fulfill in a classroom, how it would interact with students, teachers and its environment, what it could look like, and what would specifically differ from other types of social robots used in education. In the workshops we used online collaboration tools, supported by a sketch artist, to help envision a social drone in a classroom. The results revealed several design implications for the roles and capabilities of a social drone, in addition to promising research directions for the development and design in the novel area of drones in education.

KEYWORDS

social drone, education, human drone interaction, robot design, remote design workshop, robots in education

1 Introduction

Advances in Human-Robot Interaction (HRI) have recently opened up for the rising research field of Human-Drone Interaction (HDI). The field generally started by investigating novel interaction approaches such as defining visual representations of a drone Szafir et al. (2015), designing ways for motion control [Obaid et al. (2016a); Walker et al. (2018)], exploring social body motions Cauchard et al. (2016), or defining interpersonal spaces Yeh et al. (2017). In parallel, researchers have looked at utilizing drones in several application domains [see Obaid et al. (2020a)], such as entertainment Rubens et al. (2020), sports [Romanowski et al. (2017); Mueller and Muirhead (2015)], domestic companions Karjalainen et al. (2017), local services [Obaid et al. (2015b); Knierim et al. (2018)], videography Chen et al. (2018), art Kim and Landay (2018), and more. A recent review by Baytas et al. (2019) on designing drones, suggests that drone application domains that target domestic-human environments can be defined as "*social drones*". Based on a more recent HDI survey by Tezza and Andujar (2019), it is foreseeable

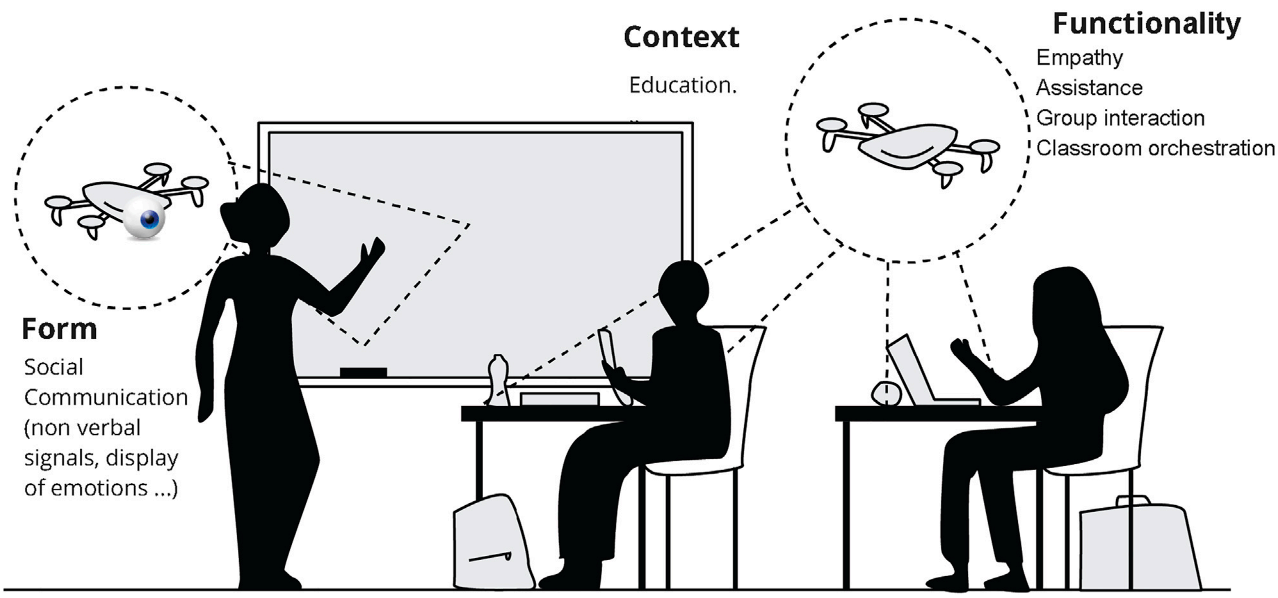


FIGURE 1
An illustration of social drones in a classroom, with examples of instantiating the social form and social function in an educational context [extending the definition by Hegel et al. (2009)].

that drones will become a ubiquitous technology deployed in many new application domains within our society, but they have not yet been investigated. In their survey, it is suggested that one way to move forward is to gauge research efforts into activities that will elicit design implications for the different societal application areas, enabling a better understanding and acceptance when utilizing drones in our society.

Extrapolating from the social HRI field, a large body of research efforts have been put towards the application domain of education Johal (2020). One of the aims is to introduce novel ways to support teachers in classroom environments Belpaeme et al. (2018), thus enhancing the students' learning capacity. In this context, ongoing HRI research suggests that assistant classroom robots is a preferred approach Ahmad et al. (2016). However, to the best of our knowledge, the application domain of social drones (or flying social robots) in education has not been researched yet. Therefore, we believe this is an opportunity to explore the design space and implications of drones in an educational context, in particular looking at drones that support the classroom environment with no intention of replacing the teacher (see Figure 1). We do this by taking a novel first step into exploring the educational drones' design space and contributing the following to the HDI community:

- Conduct novel research to understand how social drones can be utilized to support the classroom environment, teachers and students.
- Create a user-centered design method to elicit design implications on four main design themes, thus providing insights on social drones in education and future research directions.
- Identify the implications and lessons to be learned from the online and remote design workshops.

2 Related work

In this section, we highlight related research to demonstrate the need to establish a design space for a novel application area of social drones in education. The section is divided into three parts: social robots in education, user-centered robot design, and related work towards our approach to drones in education.

2.1 Social robots in education

Since the establishment of the Human-Robot Interaction (HRI) field, a considerable research body has contributed to investigating social robots in education [Mubin et al. (2013); Belpaeme et al. (2018); Johal (2020)]. Initially, educational robotic agents originated from research on virtual agents that aimed at enhancing the learning environment of students [Johnson et al. (2000); Yılmaz and Kılıç-Çakmak (2012)]. Thereafter, the physical embodiment of a robot in a classroom environment has attracted the attention of researchers, strengthening several dimensions in a learning classroom setup [Saerbeck et al. (2010); Köse et al. (2015); Kennedy et al. (2015)]. A recent review by Belpaeme et al. (2018) highlights the benefits of having a physically embodied social robot that has a tutoring role in a classroom. One of the benefits is the ability to foster engagement, creating a positive learning experience for students. In addition, in their review Belpaeme et al. (2018) demonstrated that HRI literature used a wide variety of robot appearances; pointing out that almost all of them had social attributes and features (i.e. humanoid features such as head, eyes, mouth, arms or legs). Moreover, in another review on robots used in education, Mubin et al. (2013) showed that the role of a classroom robot is generally seen as an assistant or a tutor supporting the teacher and students. To this end, the aforementioned literature reviews suggest that social robots in education are likely to be autonomous in their movements and will depict an assistant to a teacher role in a classroom environment.

In the context of social robots in education, many researchers have worked on investigating interactions using different robotic platforms that are already available in the market, such as the popular NAO robot [Johal (2020); Amirova et al. (2021)]. For example, Serholt et al. (2014) deployed a NAO robot to investigate the school children's response to a robotic tutor compared with a human tutor while giving instructions to accomplish a task. Another example is using the NAO robot Obaid et al. (2018a) to study the development of empathetic robotic capabilities in educational settings. While such studies hold a significant value in the development and understanding of educational robots in classroom environments, most can be considered to be taking a robot (device)-centric approach to understand robots in classroom settings; thus, revealing few insights about the users' views of a what a robotic agent should entail from a user-centered design (UCD) prospective. In the review by Johal (2020), it is also noted that research into the educational use of social robots in adult higher education has been almost unexplored.

2.2 User-centered robot design

Users' contributions in the design process of a robot can help fast track their acceptance and usefulness Reich-Stiebert et al. (2019b). Recently, several researchers have employed and developed new UCD approaches in the design and development of robots [Barendregt et al. (2020); Reich-Stiebert et al. (2019a); Leong and Johnston (2016); Sabanovic et al. (2015)]. Focusing on the domestic robots, the work by Lee et al. (2012b) inspired other researchers to involve users in the design of assistant robots in a classroom environment [Obaid et al. (2015a; 2016b; 2018b)]. Although their work was not focused on robots in education, Lee et al. (2012b) suggested an innovative approach to envision a domestic robot by utilizing sketches and drawings along with four main design themes: the look and feel, interaction modalities, social role, and desired tasks. Later, Obaid et al. (2015a) used a similar UCD approach to investigate how adult interaction designers and school children would envision a robot as an assistant to a teacher in a classroom. In their study, 24 interaction design students and 29 children took part in focus group sessions to draw, describe, and discuss the creation of robot designs based on the aforementioned themes suggested by Lee et al. (2012b). Their results revealed interesting insights into the clear differences between the adult interaction design students' and children's views. For example, children wanted their robot to have a human-like form that included robotic features, but adult designers envisioned a cute animal-like robot in a classroom space. Thereafter, Obaid et al. (2016b) developed a robotic design toolkit (Robo2Box) aiming to support children's involvement in the design of their classroom robot. The work presented by Obaid et al. (2018b) suggests that social features were envisioned by children. In addition, some children expressed the preference of having a robot with flying capabilities in a classroom.

2.3 Human-drone interaction and social drones

One direction in recent HRI research is increasing activity in investigating flying robots (drones), thus creating a whole sub-field on Human-Drone Interaction (HDI). In a recent HDI survey by Tezza and Andujar (2019), research is currently focusing on (1) exploring ways of HDI communications, (2) identifying suitable interaction modalities with drones, (3) investigating human-drone social behaviours (e.g., proxemics), and (4) introducing novel application areas and use-cases. Generally, the two first items are directed towards investigating novel modalities of interaction between humans and drones and the last items investigate novel application domains for these interactions.

For example, studies ranged from exploring ways to navigate/control a drone using body gestures [Obaid et al. (2016a); Cauchard et al. (2015); Ng and Sharlin (2011)], to utilising visual representations held/projected by drones [Scheible and Funk (2016); Schneegass et al. (2014); Romanowski et al. (2017)].

In the case of social drones, however, there is some specificity that will apply. In this section, we propose to define the characteristic of social drones grounding the definition of social robots and highlighting the specificity

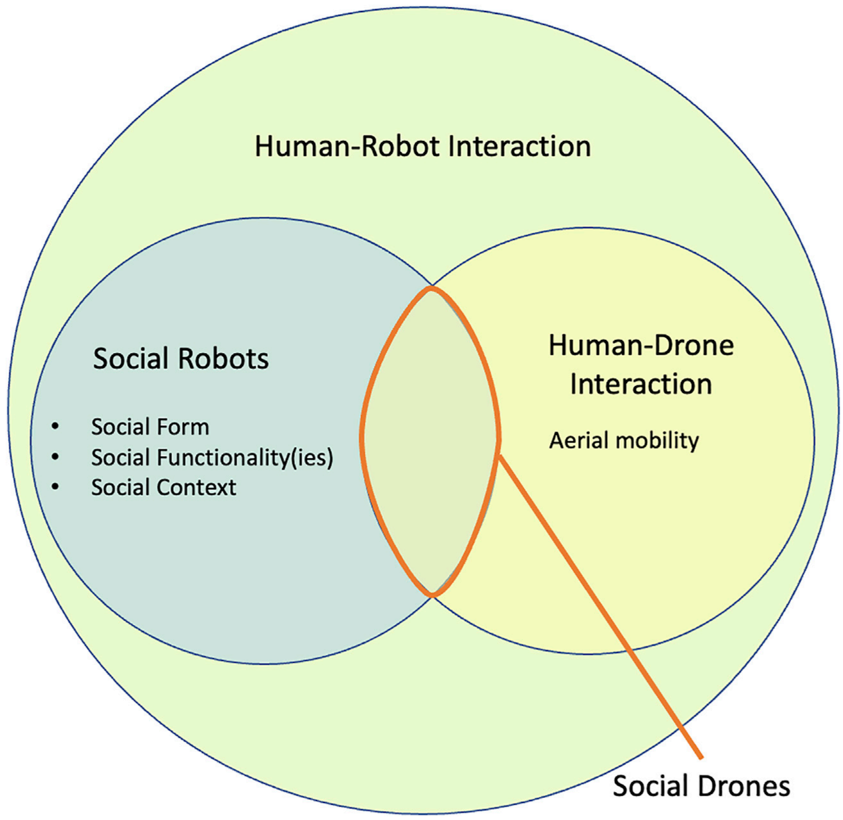


FIGURE 2
Overview of the research areas linked with social drones.

of social drones. The literature on the design of social drones was recently reviewed by Baytas et al. (2019). In their review, they gave a summary of three main design categories and twelve design aspects that were identified from related literature. The design aspects were put into perspectives in either Drone Design concerns (lighting and displays, proxemics, sound, appeal, control methods, form, flight) or Human-Centered concerns (ergonomics, intuitive control, perceived social role, tactility perception, and intuitive comprehension). While this review is the first focusing on social drones, the authors only provided a short definition, mainly relying on the context of deployment of the drone "*we submit that an autonomous embodied agent in an inhabited space can be similarly described as social.*" While important, social context alone is insufficient to qualify a drone as social.

In Human-Robot Interaction, Hegel et al. (2009) proposed to define social robots as a composition of robot and social interface, with the social interface referring to all the designed features by which a user judges the robot as having social capabilities being: 1) its form, 2) its functionality, which should be exercised in a socially appropriate manner, and 3) its context of use. This definition of social robots can be further defined thus:

Form: All aspects related to the robot communication: its appearance, its socially expressive capabilities (i.e., facial expression, display of emotions, expressive navigation, other types of non-verbal cues)
Function: All aspects which compute any artificial social behaviour of a social robot: BDI reasoning, joint attention mechanisms, theory of mind, affective recognition, and others
Context: All aspects linked to the knowledge of specific applications domains, in terms of social norms and social roles

While presented distinctively, these aspects are obviously interrelated. For example, social functions and social forms would be aligned to meet a user's expectations (e.g., one would expect that a robot with a mouth could speak).

Going back to the HDI research context, we can now distinguish the social drone as a special case of social robot thanks to its aerial motion capabilities. Looking at the recent literature review by Obaid et al. (2020a), we hence see that researchers are investing these aspects. For example, Herdel

TABLE 1 Sessions and participants' information (P#: Participant ID, F: Female, M: Male, D: Design, E: Engineering), All the teaching experience reported was university level.

Session ID	P#	Grad	Gender	Background	Teaching experience
S01	P01	Bachelor	M	E	
	P02	PhD	F	D	Y
S02	P03	PhD	F	D	Y
	P04	Master	F	D	
	P05	PhD	F	D	Y
	P06	PhD	F	D	Y
S03	P07	Master	F	E	
	P08	PhD	F	E	
S04	P09	PhD	F	E	Y
	P10	PhD	M	E & D	Y
	P11	PhD	F	D	Y
S05	P12	PhD	F	D	Y
	P13	Bachelor	F	D	
	P14	PhD	M	E	
S06	P15	Bachelor	F	D	
	P16	Bachelor	F	D	
	P17	Master	M	E	
S07	P18	Master	F	D	
	P19	Bachelor	F	E & D	
	P20	Master	M	E	

et al. (2021) propose to investigate how facial expressions and emotions could be rendered on a face mounted on a social drone (Form). In another example, Karjalainen et al. (2017) investigate how social drones could be used for companionship in the home environment, and what functionality and role these drones would have. Figure 2 summarises the characteristics of social drones in relation to HDI (e.g., interested in interactions between humans in flying robots) and general social robotics.

To summarize, looking at the importance of social robots in education within the HRI field (Section 2.1), and inspired by the domestic companion work of (Karjalainen et al. (2017)) and other related work (Sections 2.2, 2.3), we apply a user-centered method to take the initial steps towards exploring a novel design space on social drones in education. In the following section, we describe our method and the UCD approach in detail.

In particular, we aim to establish how the role envisioned for social drones differs from that of classical social robots in education; what features of the social drones (e.g. aerial motion, bird-view) are envisioned to be useful; and finally, what threats and risks are envisioned in the case of social drones in classrooms.

3 Methods

To investigate social drones' assistance in educational contexts, a set of seven design workshops were conducted remotely, each with 2-4 participants, 2 facilitators, and 1 sketch artist. The study was approved by the institute's ethics committee and all participants were recruited on a voluntary basis. Their consent was obtained through an online consent form prior to the study. The workshop structure consisted of two steps and the duration of each workshop was 1.5 h on average. A few days after the workshop, a post-workshop questionnaire was sent to participants to evaluate the techniques and tools used (i.e., sketch artist's support and online tools).

The essence of this study was to take a learner-centered design (LCD) approach, in order to gauge the learners' environment of an educational drone. We specifically focused on Higher-Education student who are often less researched as a target group for social robots in education (see Johal (2020)).

3.1 Participants

An announcement for the study was made via different online channels, including social network accounts and email lists to students at the authors' institutions. In total 20 participants (15F and 5M) responded to the call and attended the workshops. The workshop announcement was an open call, so there were no criteria for participant selection, however, all the respondents were either university

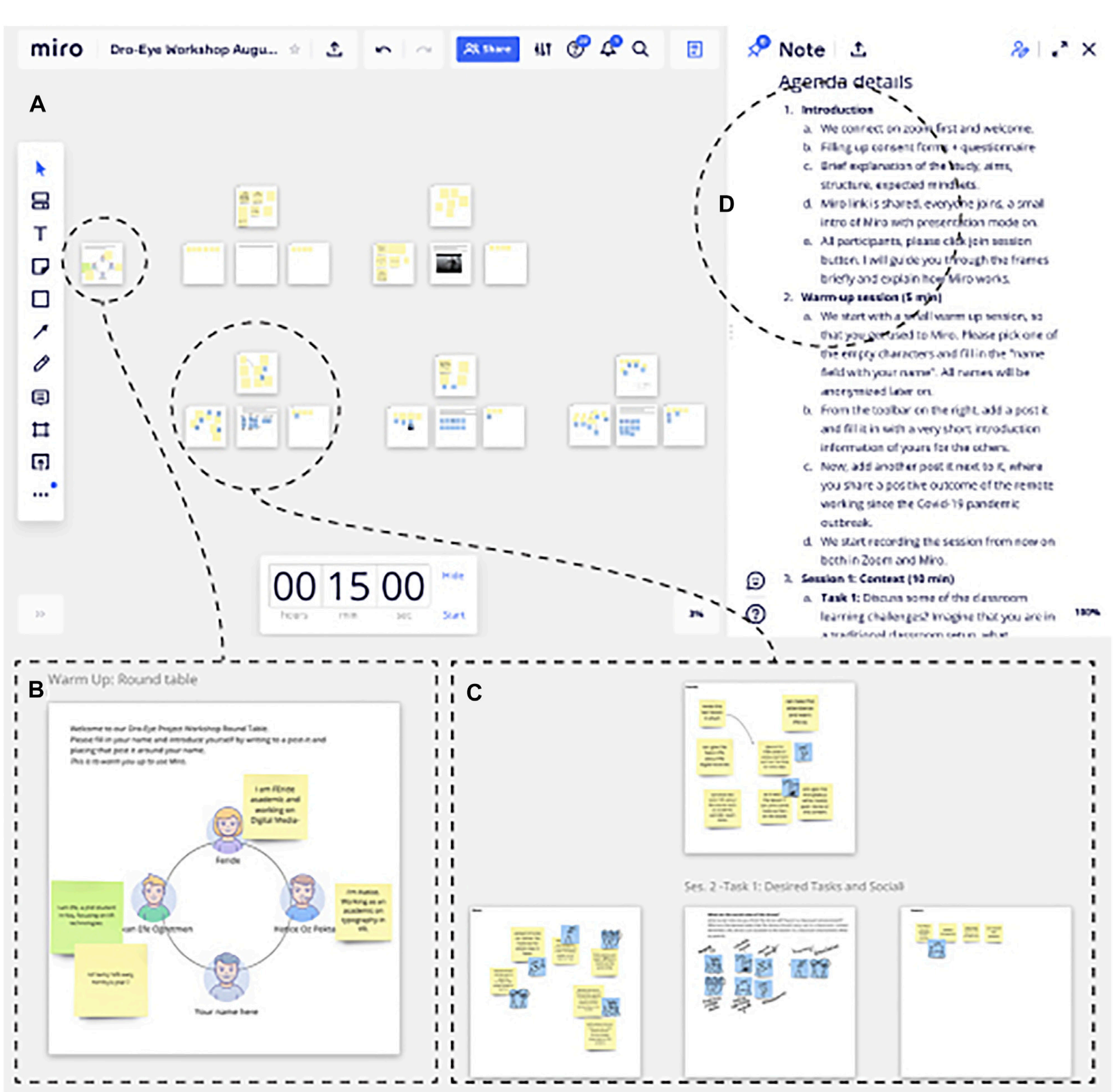


FIGURE 3
Scene from a Miro session. **(A)** The Miro canvas contained five distinct board groups. **(B)** Orientation board and avatars. **(C)** Each task consisted of one central board and two to four surrounding boards for participants. Each participant had a dedicated seat around the central board, resembling a round-table setting. **(D)** Easy to access workshop agenda for participants' and facilitator's reference.

students or graduates (students: three BA, three MA, seven PhD; graduates: two BA, two MA, three PhD). Of the 20 participants eight had teaching experience, while six were specialized in education-related research, four of whom worked particularly on "education for children". Finally, 10 out of 20 participants had engineering-related backgrounds and 12 out of 20 had design-related backgrounds (two had both backgrounds). One workshop was conducted with four participants, four workshops with three participants, and two workshops with two participants. Table 1 summarises the important demographic information of participants for all the workshop sessions.

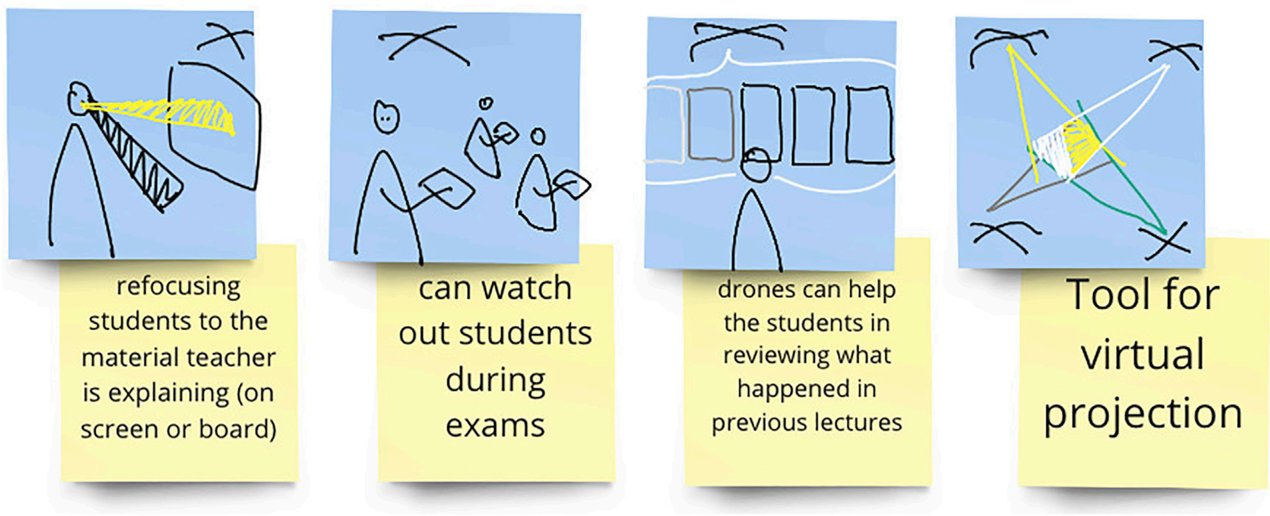


FIGURE 4
An example of the drawings made by the sketch artist to visualize participants' written ideas during the fourth workshop. The hovering X shape illustrates a drone. The shape representing the drone was suggested by the sketch artist during the pilot study, to allow the visualization of drones in the quickest way possible.

3.2 Study settings

All workshops took place via two online platforms concurrently: Zoom[1], the video conferencing platform, and Miro[2]. The latter is an online collaboration platform that allows users to share their ideas and comment on shared artifacts on an infinite canvas using sticky notes, drawing tools, and customizable boards (see Figure 3). Besides collaboration, Miro also offers facilitation features such as a timer, an easy-to-access workshop agenda, and directing users' attention to a specific spot when needed using the visible mouse mode. First, a Miro canvas template was created and used in a pilot study with one participant to evaluate the ease and the flow of the study; thereafter, necessary adjustments were made to revise the workshop structure and the Miro template. The revised template setup was duplicated and used for each workshop. While everyone communicated through Zoom, Miro was used for generating ideas and having discussions on the matter during the workshops. Zoom sessions were audio-recorded, whereas Miro sessions were saved as PDF files and participants' ideas were exported to a spreadsheet for analysis. One researcher recorded data and managed the technical aspects of the workshop without interacting with the participants, and one researcher undertook the facilitation, guiding the participants through all tasks and managing the discussions as well as keeping track of the time. The reason to use Zoom instead of Miro's audio communication was to avoid connection problems.

Each session consisted of two steps, each having three tasks, as explained in the below sections. In each task, participants used personal boards assigned to themselves around a central board, resembling a physical round-table setting (see Figure 3). The central board was used by the sketch artist to visualize participants' ideas while they talked and wrote down their ideas related to each of the tasks. When a task was completed, participants moved to the next round-table to work on the next task. Each participant's personal board contained three empty sticky notes for them to start each task. They were free to add more but had to at least fill one.

3.3 Step one: Context building

The aim of the first step was to familiarize the participants with Miro and the topic of the workshop. It consisted of three tasks:

3.3.1 Task 1: Orientation

Participants were welcomed via an orientation board, where they were asked to choose an avatar, write their name next to it and introduce themselves using sticky notes next to their chosen avatars. After each participant figured out how to use the Miro basic tools needed for the rest of the workshop and had met each other, everyone moved to the next task. For the rest of the workshop, participants had the same position around the board

1 https://zoom.us/
2 https://miro.com

based on the one they picked in this first task, as if they were sitting around a table.

3.3.2 Task 2: Recall a learning challenge

As an introduction to the topic of the study, participants were asked to write down the challenges they face (or had faced) in physical education settings. Each participant presented their sticky notes, and discussed the challenges and issues they had experienced in classrooms. We noticed here that the participants referred mainly to issues encountered in their higher education context.

3.3.3 Task 3: Presentations of current drone applications and capabilities

The general public shapes their conceptual understanding of drones (as a rising technology) from representations exhibited in the mainstream media (e.g., movies); as illustrated by Aydin (2019). This meant that our participants arrived at the study with some previous knowledge of drones from the different media channels. Aydin (2019) also states that the general public is not aware of the majority of applications that drones can hold, giving us an incentive to educate our participants on a broader understanding of the *social drone* space. To achieve this, we used a priming video[3]. After the discussion on the challenges in educational contexts, all participants were briefly introduced to the many current applications of drones by the facilitator, and then watched a YouTube video on social drones created by the MagicLab (2016), illustrating how drones, individually and in groups, could interact with a user, and what their current interaction capabilities are.

3.4 Step two: Brainwriting on social drones in education

In this step, the aim was to investigate how participants imagine drones being used in educational settings. Inspired by the work presented by Obaid et al. (2015a), we adopted similar themes to reveal design aspects on social drones in education. Therefore, this step consisted of three theme-based tasks: desired tasks and social roles, interaction modalities, and the look and feel of social drones. Furthermore, for this step, an adapted brainwriting Rohrbach (1969) technique was used, in which participants were asked to individually note down their ideas on sticky notes on their boards before exchanging those ideas in a group discussion. Concurrently with the brainwriting activity, the sketch artist made drawings of participants' ideas on the central board. Each idea was captured with drawings on separate sticky notes, and those drawings were improved based on the discussion among participants and on their feedback (see Figure 4). Participants were also asked to match the drawings with their own ideas during the discussion; in order to help others to understand and to facilitate the discussion further.

3.4.1 Task 4: Desired tasks and social roles of social drones in education

For this task, the facilitator asked the following questions to the participants and set up a timer for 5 minutes for brainwriting:

- What are the social roles of a drone?
- What social roles do you think the drone will have in a classroom environment?
- What are the desired tasks that the drone should carry out in a classroom context?

Participants were reminded that the drone should be imagined as an assistant supporting a classroom environment, possibly addressing some of the learning challenges they elicited in Step One, Task 2. After the brainwriting activity, the drones' possible social roles and tasks in educational settings were discussed.

3.4.2 Task 5: Interaction modalities

In this task, the goal was to identify how participants would envision interacting with a drone in an educational setting. They were asked to generate ideas based on the following questions:

- What interaction modalities should this drone have?
- How do you envision your drone interacting with students, with teachers, and with the environment?

The time given for brainwriting in this task was again 5 min.

3.4.3 Task 6: Look and feel

Based on the discussion on previous tasks, the facilitator asked the participants to envision the look and feel of social drones. Within 5 minutes, the participants were asked to think about and note down their ideas on the following questions:

- What do you imagine the drone to look like?
- What would its shape, color, size look like?
- Are there any additional functions or parts that the drone should have that would help with its social role?

3.5 Post-workshop questionnaire

A week after each workshop, a questionnaire was sent to the participants to collect their feedback regarding the workshop's form and content. Each session's questionnaire started with the pdf export of their workshop discussion boards and participants were asked if they wanted to add anything. The rest of the

3 Video used for priming: https://youtu.be/B4xtsH6pzoM

TABLE 2 Coding scheme used for the quantitative analysis with possible codes separated by ";" for each category.

Main category	Possible codes
Social roles	Teaching assistant; Butler; Moderator (i.e. between students or between teacher and students); Environment Manager (i.e. classoom facility); Companion; Entertainer; Special Needs Caregiver
Interaction modalities: Social cues input	Gaze tracking; Movement tracking; Face/Object recognition; Video/Image capturing; Speech; Environment sensors; Controller (UI or tangible); Gesture control
Interaction modalities: Social cues output	Gaze; Speech; Non-verbal sound; Lights; Displays; Body color; Body language; Writing or texting; Drawing; Tactile feedback
Drone size (diameter)	Small (< 20 cm); Medium ($\geq 20cm$ and $\leq 35cm$); Large ($> 35cm$)
Desired tasks	See Table 3 for details on the codes

questionnaire focused on collecting participants' feedback on the form used to run the workshops (i.e., online tools and the sketch artist's support). The output of the questionnaire and participant's feedback are presented and discussed in Section 5.4.

3.6 Data collection

All the sticky notes were extracted from Miro and parsed into a table for annotation. The first reading of all the sticky notes allowed us to clean and remove empty ones. The audio recording was also used when the idea written on the sticky notes was not clear. The recording was then transcribed to complement the information on the sticky notes. The analysis process was mainly based on annotating the data logs for the participants from each of the workshop sessions. In this section, we describe the derived coding scheme and the annotated coding process.

3.6.1 Coding scheme

We utilised a deductive categorisation approach [see Bengtsson (2016)] to define the coding categories, mainly based on previous literature, the coders' knowledge, and experience in the HRI field. The coding scheme was built to address our three main research questions. When constructing the coding grid, we referred to the work by Lee et al. (2012b) (i.e., Role, Look and Feel, Desired Tasks and Interaction Modalities), and Yanco and Drury (2004)'s taxonomy to characterize the interaction context (i.e., interaction roles, task type), to define the level of granularity, and Obaid et al. (2015a)'s list of robotic attributes that helped to draft the interaction and look and feel codes (i.e., size, sensors). As a first step, the coders started with the main coding categories adopted from the design themes presented in previous research [Yanco and Drury (2004); Lee et al. (2012b); Obaid et al. (2015a)]. The main themes consisted of the social role, desired tasks, interaction modalities, and look and feel. The coders then started discussing the categories, and iteratively defined coding options together to derive an initial coding scheme. Thereafter, each coder used the initial scheme to annotate a sample data set of 55 sticky notes, allowing the coders to discuss and amend any further code categories to the scheme. The final coding scheme contained the main categories listed in Table 2.

3.6.2 Coding process

The groups generated various notes during the session. The process started with four coders analysing a whole session, 10% of the data, to deduce an inter-rater reliability check. Each coder separately analysed the data logs of one entire session, composed of 55 sticky notes. Thereafter, an inter-rater reliability was computed via a Fleiss measure (Fleiss (1971)-suitable for more than two raters for categorical ratings). On average, across all of the dimensions coded, the inter-rater reliability resulted in a Fleiss' kappa of 0.69 (SD = 0.13), which is considered to be a substantial agreement. The coders discussed some of the categories further to eliminate any misunderstanding or ambiguities throughout the process. The annotations of the full dataset then commenced, where each of the coders was assigned a set of coding categories to annotate across all of the seven workshops.

4 Analysis and results

This section presents the results of the analysis of the ideas captured by the sticky notes and the sketches across seven workshops (20 participants). All of the participants generated a total of 463 sticky notes. In order to report the results, we propose to articulate them following the design themes proposed by Lee et al. (2012a).

4.1 Social roles and desired tasks

Participants (P) were invited to envision what social roles would be desirable for a drone in education to support and assist in a classroom setting. Out of the 498 sticky notes generated during the seven sessions, 100 referred to a social role for the drone in the classroom.

TABLE 3 Coding scheme used for the desired tasks for the social drone in classroom.

Task	Occurrence
Managing participation (i.e., answering small questions, grouping students)	34
Doing chores (i.e., cleaning, transferring items, bringing food)	30
Proctoring exams/taking attendance	10
Enriching demonstration and visualization	8
Managing students' attention/focus	6
Ensuring safety/assisting people in case of emergency	6
Ensuring physical settings and arrangements	6
Ensuring everyone understands, hears the teacher, understands the language	3
Informing students about past/current/future class content	3
Assisting musical/body exercise	2
Recording class/course content/board	2
Enriching teacher's gestures	1
Broadcasting teacher's lecture in multiple classes	1
Keeping track of time	1
Helping teachers to evaluate themselves	1
Understanding students' mood	1

The results show two main roles that were predominantly mentioned by the participants: Teaching Assistant (mentioned 54 times), and Butler (22). Other roles mentioned were Moderator (8), Companion (7) and Environment Manager (6).

While Teacher Assistant/Tutor is quite a commonly found role for social robots in education Belpaeme et al. (2018), the task envisioned for social drones as teacher assistant differs somewhat from that of classical social robots in education.

Furthermore, the role of Butler often cited by our participants has mainly heretofore been in the literature about social robots in a home context Karjalainen et al. (2017) and not educational contexts. The ease of movement of drones could explain this as many butler tasks consist of moving objects around the room.

Table 3 shows the envisioned set of tasks for the social drone. Out of the 498 sticky notes, 122 mentioned a task that could be carried by the social drone. While the set is very varied (we counted 17 different types of tasks), two stand out by the frequency at which they were mentioned by the participants: Managing participation (mentioned 30 times) and *Doing chores* (mentioned 21 times). These tasks are in line with the principal desired social roles discussed before, respectively Teaching Assistant and Butler[4].

We notice that these tasks differ from classical tasks performed by social drones in other contexts, more often found to be suitable for navigation, well-being, and companionship (Obaid et al. (2020a)). But they also differ from classical tasks attributed to other types of social robots in education which often offer individual assistance Johal (2020) rather than classroom orchestration functions.

In terms of the envisioned target users for the social drones[5], we notice that the participants mainly mentioned tasks involving individual interactions between the social drone and one student (26 times): "*[It] can help students taking notes when the teacher moved on, it can [. . .] explain any point students missed/didn't understand*" (P16). They also thought the drones could interact with the whole classroom (11 mentions) by broadcasting information, being used as a tool to do physical demonstrations, assessing and managing the well-being of students at the classroom level: "*Classroom environment can be modified to enable a more productive climate (heat, seating arrangement, etc)*". Finally, some participants thought the social drone could be useful for group work (6 mentions): "*During teamwork it can guide students or participate as well like a leading team member*" (P16).

4.2 Interaction modalities and communication

In *Task 5*, participants were asked to imagine how users could interact with a social drone. We categorised their comments into two types: 1) Social Cues Input-describing how the drone

4 see the dashboard for the full list: https://apps.streamlitusercontent.com/wafajohal/socialdroneeducation/dev/streamlit_app.py/+/#task-analysis

5 see the dashboard for the full list: https://apps.streamlitusercontent.com/wafajohal/socialdroneeducation/dev/streamlit_app.py/+/#target-user

perceives its environment and the users, and 2) Social Cues Output-describing how the drones communicate with the users. The analysis shows that participants often mentioned multiple modalities for the drone to perceive its environment and the users, among which the most mentioned were: Speech/Audio (13), Environment sensors (i.e. temperature, light) (10), Gesture (10), Movement Tracking (9), and Face or Object Tracking (7)[6]. We note here that it is interesting that the participants didn't think about the fact that the drone's propeller might be too noisy for microphones to be used. We also note that touch was not envisioned and that the modality of interaction proposed by the participants could be implemented for public space proxemic distance see Mumm and Mutlu (2011).

In terms of social drone expressivity, participants mainly mentioned non-verbal communication cues (only 3 sticky notes mentioned Speech and 2 Non-verbal sound). Body Language was the most mentioned modality (8). For instance, participants thought of using different types of motion to convey messages: "*fast-short movements* vs. *slow continuous movements*" (P12), or to include gestures such as "*nodding*" (P20 and P03). Several participants mentioned lights (4) or body changing colors (3) as a way for the drone to express its mood "*RGB color codes: Green is positive, Red is negative [and] Blue is uncertain*" (P20) or to signal learning phases to the whole classroom: "*The color of the light of the drone might indicate something such as its the time of the lecture for questions and discussion*" (P18). The sound of the drone was often considered to be distracting and only two participants mentioned it as a possible way for the drone to communicate. Three participants also mentioned the possible use of a projector embedded in the drone: "*projection (light, picture)*" (P07). Nearly all the interactions mentioned by the participants were spatially collocated (84), and only three mentioned remote interactions with the drone. This has implications for the drone's appearance (i.e., in terms of safety and discretion), discussed below.

4.3 Look and feel

Nineteen out of the 20 participants mentioned the drone's size in their sticky notes. To code for the drone size, we decided to split the 19 sticky notes into three categories relative to the size of a regular commercial drone such as the Phantom Dj3 (see Table 2). Analyzing the results, we find that participants mainly mentioned a small hand-sized social drone (10 out of 19 sticky notes). It was often justified by the fact that the drones needed to appear safe and to be discreet. Only two sticky notes mentioned a large drone, anticipating that the drone needed to be big enough to contain and carry objects in the classroom or to arrange the table and chairs after the class.

Related to the drone's size and shape, three participants mentioned that the drone could change shape as a way to illustrate concepts and enrich presentations:"*take the shape of the Eiffel Tower during a lesson on France*" (P08), or to be transportable: "*folding/unfolding for saving space and being able to use in different sizes*" (P06). Further, on the drone's shape, three participants suggested that drones should look appealing to children. To that end, cartoon-like, animal-like: "*butterfly*" (P08); "*similar to bird or animal*" (P10); or object-like such as a "*balloon*" (P14) or a "*toy*" (P06) drones were mentioned by several participants. Related to child-friendliness, participants also suggested colorful and easy-to-distinguish drones: "*very colorful like a flying insect*" (P03).

Another aspect related to the look and feel of the drones was the potential distraction that the drones may cause. Seven participants explicitly addressed that the drones should be as silent as possible. Possible solutions suggested for making drones silent were covering or removing the propellers (P19, P10) and using noise-cancelling (P20). One participant mentioned that the drone should be "*silently lurking around*" (P05) so as not to interrupt people in the class. Besides the auditory distractions, five participants were concerned about visual distraction, and suggested the following to render drones less distracting: "*transparent drone*" (P21), "*seamless: change the color according to the environment*" (P11), "*not a vibrant color, plain physical features*" (P10) and "*ghost-like*" (P10).

Non-threatening drones emerged as a need in the classrooms both for children and older students. Participants mentioned drones to be physically harmless and with a friendly appeal. Related to physical safety, soft materials (P11) were suggested. Another suggestion for harmless drones was that "*lights should not be harsh or blinding*" (P20).

In addition to the features related to look and feel explained above, lightweight (P11), color-changing (P10, P02) (depending on e.g., students' mood (P06, P04), environment (P08), preferred visibility (P12), performed behaviour (P20), social role (P13), activity status (P09) and class subject (P08)), smooth (P03) and modular (P06) drones were suggested.

5 Discussion

The main aim of this study is to explore how social drones would be utilised in a classroom setting, in terms of social roles, tasks, interaction modalities, and appearance. Here in this section, we will discuss implications related to the design of social drones in a classroom and the methodological implications associated with the remote design workshops.

6 see the dashboard for the full list: https://apps.streamlitusercontent.com/wafajohal/socialdroneeducation/dev/streamlit_app.py/+/#social-cues-input

5.1 Implications for social role

Keeping in mind their learning challenges, several roles were envisioned by participants in the study, but two main roles stood out: a Teaching Assistant (TA) and a Butler. A TA was by far the most mentioned role for the social drone in the classroom. Participants thought of the TA as being able to bridge the communication between the teacher and the students, either by using other media (recording, broadcasting, enriching the demonstrations and visualisations) or through gestures ("*attract the attention of teacher*" [P03]). Similar to a regular classroom TA, the social drones could be asked to gather questions, answer some of them directly, manage group discussion, and balance students' participation. Low Engagement and lack of focus during the class were discussed to be important challenges by several participants. While several participants thought that the TA could be more accessible and engaging than the teacher, they also expressed concern about the drone being distracting. As a result, the look of the drone incorporates attributes such as small, invisible, and quiet: "*size-not so big, not to distract students so much*" (P16), "*seamless[. . .]*" (P11) "*transparent drone*" (P21), and "*silent as possible as it can be*" (P10). To summarise, the drone TA is seen as a way to discreetly bridge communication between students and teachers through physical movement.

A Butler was the second most mentioned role, and this is an unexpected role when applied to social robots in education (Belpaeme et al. (2018)). The butler was seen as an assistant in charge of students' well-being in the classroom, that "*brings food when people are hungry*" (P18), "*handing out/collecting tools, papers*" (P19), "*clean[ing] any left trash[, and] adjusting seats and tables after class is done*" (P01). Because of its ability to easily move in space, the butler drone would be expected to carry objects, and hence was thought of as a larger drone, sometimes with "*extensible arms*" (P01) or a "*small storage (for distributing things)*" (P13). Here also, one aspect participants thought of as important was a safe appearance: "*appear harmless*" (P03), "*it must be made of a soft material due to safety reasons*" (P11), and "*not threatening*" (P20). This new butler role brings new scenario opportunities for social robots in the classroom, closer to what is commonly found in home settings.

5.2 Implications for the communication capabilities

The interaction modalities that support the social capabilities of a drone were largely discussed by participants, especially when envisioning how the drone could sense users and its environment. We noticed that participants were more talkative about the social inputs and sensors that the drone could have (55 sticky notes on input sensing capabilities) compared to its ability to express itself (only 28 sticky notes on the output capabilities). This could be a result of the fact that 10 out of 20 participants have an engineering technical background. This has also been seen in previous HRI research, in which interaction with humanoid robots was clearly influenced by the technical or non-technical background of the participants Obaid et al. (2014).

Some participants also thought the drone could be in the classroom to monitor student attendance or exams, hence to "*detect humans*" (P19), or to "*give a warning if it detects suspicious behaviour during exams*" (P13). The well-being of students in the classroom was often discussed with two aspects: the students' mood (affective state and attention) and environmental factors (temperature and light adjustments). Several participants agreed that the drone could be used to monitor and adapt to a student's mood: "*Biosensors (Understanding the mood of the user)*" (P02), "*mood detection*" (P18); and to the environment: "*adjust the physical environment for best learning conditions (light, temperature, screen size, etc)*" (P16) and "*[the] classroom environment can be modified to enable a more productive climate (heat, seating arrangement, etc)*" (P05). Social capabilities were attributed to the drone to sense and interact with students similarly to other social robots in education (Belpaeme et al. (2018); Johal (2020)). Additionally, we found that abilities to sense the whole classroom were proposed for social drones by participants which is quite specific to the drone's ability to adopt a bird's-eye position and see the entire class (see Section 6.1.4). Having said that, improving social expressivity of social drone is needed as they lack anthropomorphic features used by other social robots in education [e.g., NAO Amirova et al. (2021)].

5.3 Implications for the appearance

Some participants thought of the social drone as a *versatile* agent that could change appearance and/or change role: "*modular [drone]*" (P06), "*different colors [could] indicate different social roles (janitor [in] red, helper [in] green, assistants [in] orange etc.)*" (P13), "*changeable to the [curricular] subject and to the environment*" (P08). In addition, two aspects of appearance that predominated the participants' comments were safety and discretion. These two last aspects are very specific to social drones. Indeed, noise and visual distraction could be a threat for students' learning. We discuss this aspect further in Section 6.1.1.

5.4 Implications for remote design workshops

In this section, we discuss lessons learned from our experience with the online combination of using Zoom, Miro, and sketches employed in a design workshop for drones in education.

5.4.1 Designing online around a virtual table

Miro has provided an online collaboration platform for participants and helped in materializing their thoughts. The 18 participants that responded to the post-workshop questionnaire agreed that it was easy to share their ideas using Miro (4.28 on the 5 point Likert scale). Participants each had their own boards that were positioned as if they were sitting around a table, and when it was time for group discussions, they used the group board in the middle. This setup allowed participants to have individual thinking time when needed and orient joint attention during presentations and discussions. Miro has enabled us to overcome certain challenges encountered in co-located physical workshops, such as getting distracted by others or having difficulties in presenting ideas placed on sticky notes due to the distance among participants. Moreover, this online-remote workshop helped us bring together participants, facilitators, and the sketch artist from all around the world (Turkey, Australia, Sweden and Portugal).

Along with Miro, Zoom was used for the introduction of the workshop, audio communication and recording during the Miro sessions. The decision of leaving the video on or off was left to the participants but it was suggested that they should prioritise their time on the Miro screen to be able to concentrate on the workshop content. Only one participant stated that it would have been better to be able to see others during discussions.

5.4.2 Designing with visual sketching support

The third component of the method we used was the sketch artist's contribution to the workshops. The idea behind employing a sketch artist was to support participants in presenting and discussing their ideas that would otherwise be written and limited to keywords. That is why the sketch artist simultaneously visualized ideas that participants noted on their boards, and the sketches were placed on the central board while participants explained their ideas in each task. Participants were encouraged to interact with the sketch artist to ask for corrections or comment on the illustrations. Again, the participants mentioned in the post-questionnaire that they were happy with the extent that the sketching artist captured their ideas (4.28 on the 5 points Likert scale). They also thought that the sketches helped them understand the others' ideas (4.17/5.00); in addition, the sketches helped them enrich their presentations (4.00/5.00). From the perspective of the sketch artist, we realized that it was challenging to visualize multiple participants' ideas simultaneously, particularly when the number of participants exceeded three. The sketch artist, in the attempt to capture a participant's idea, also often asked for clarifications in order to be able to draw a corresponding sketch. This was greatly helpful to increase discussions and have participants express their thoughts in more detail.

6 Conclusion

In this paper, we have presented a study that aims at envisioning the design of social drones in a classroom context. After running a series of seven online design workshops (20 participants) with the support of a sketch artist, we analyzed, discussed, and extracted several main design implications that can advance future research on social drones in education. To the best of our knowledge, this is the first summary of design implications on social drones in education that contributes to the HDI research and the HRI community in general. We conclude with the following list to shed light on some of the main implications found in our analysis:

1. A teaching assistant and a butler are the main social roles in the context of designing a social drone in educational settings. While the roles that came out from the workshops were relatively similar to other social robots in education [Belpaeme et al. (2018); Johal (2020)]; the findings pertaining to tasks and abilities differed. In particular, we found that the abilities for the drone to easily move in the classroom made participants think of different scenarios: handing sheets of paper, roaming in the classroom to address students' questions, and acting as a carrier pigeon to carry messages between students. Another main identified role was to be the classroom butler; as discussed in the paper, while common for home robots Lee et al. (2012b), this role has not been explored in the classroom context.
2. A social drone in education should have several interaction modalities to sense the user's behaviour, particularly the non-verbal behaviours exhibited by the user. In addition, the drone should also have the capabilities to sense its surrounding environment. These are mostly considered as input channels for the drone to make sense of its users and environment and adapt to these.
3. A social drone in education should also be able to exhibit and communicate via different interaction modalities; in particular, the drone should have expressive embodied motions as an output channel.
4. Social Drones could benefit from an adaptive design of their appearance and role. For example, different illuminating body colors could be customised to mean different roles in educational settings.

6.1 Research in social drones in education

Currently, Drones in Education can be considered to encompass several realistic scenarios despite the challenges of the technology and the associated infrastructural foundations at educational institutes. For example, drones can be utilized to be used in aerial monitoring of students on the ground, in particular in outdoor settings. Another example, is the use of drones in education as a tool or an applied instrument in educational subjects, such as science and technology. However, envisioning the use of drones as a social entity in education that can deliver and communicate with students has several research challenges that his articles tried to outlined. In this section, we demonstrate some of the challenges and associated research opportunities that entail doing research in the area of Social Drones in Education. Here we outline what HDI researchers can address in their work to help in driving the field towards future scenarios and research in HDI.

6.1.1 Challenge 1: Drones' noise in the classroom

One major issue with most of the commercial drones at the moment is the noise they make when flying Schäffer et al. (2021). Several mitigations for this issue have been proposed Watkins et al. (2020). The first one is the design of propellers that minimise noise. Other suggestions were to allocate zones and flying path to drones.

6.1.1.1 Opportunity

While this issue remains, it is difficult to envision drones being used during lectures. Researchers and drone designers need to come up with ways to further mitigate for this challenge. For example, can drone's noise be used as a way to communicate Watkins et al. (2020). Moreover, as suggested by Schäffer et al. (2021), drone noise annoyance might be lower in noisy contexts such as group work or tutorials. However, the main opportunity remains on further work to lower the noise of propellers why in motion.

6.1.2 Challenge 2: Drones lack social cues

Unlike the research literature into social robots for education that often investigate how the robot's affective capabilities could be used in learning scenarios, the mention of affective cues for social drones was nearly absent in our workshops.

6.1.2.2 Opportunity

The current form of drones is not anthropomorphic, and doesn't allow it to naturally render emotions. [Duffy (2003) makes the link between anthropomorphism and social capabilities of robots]. In order to render social cues, some previous work investigated how the drone's flight path could be used see Cauchard et al. (2016) and some very recent work even proposed to embed facial expressions or eyes on flying robots [Herdel et al. (2021); Karjalainen et al. (2017); Treurniet et al. (2019); Obaid et al. (2020b)]. Overall, developing social cues for drones to exhibit and equally perceive is a prominent opportunity for the HDI community to address.

6.1.3 Challenge 2: Drones are scary

The perception of drones as a threat was often mentioned during the workshops and this can be linked with the previous point and to common feelings about drones (Aydin (2019)).

6.1.3.1 Opportunity

During the last part of the workshop (Look and feel of the drone), participants often mentioned friendly and non-threatening look to be an important aspect. More social cues, and "warmer" material could be used to design drones that can more easily be accepted in social contexts. This leads us to further work to be done in the area of drone fabrication for HDI, which we believe is a novel area that is yet to be explored.

6.1.4 Challenge 3: Handling more than one student

In terms of the number of participants, we noticed that compared to the literature on social robots in education (Johal (2020)), there were more mentions of group drone interaction. This can be explained by the fact that drones can take this birds-eye/distanced stance, allowing them to interact with more than just one student at a time. A drone can also easily move in the classroom allowing it to be used as a novel channel of communication between students.

6.1.4.1 Opportunity

While this reduces the opportunities for individualised tutoring scenarios, it opens doors to explore a more school realistic setup in which the ratio would be one drone per classroom. It also open opportunities to expand the current paradigm in social robots for education who tend to be used for personalisation and individualised learning Johal (2020). Finally, a novel direction here can be directed towards research in Drones for Education within the CSCW arena. To the best of our knowledge, very little research has been done there addressing drones in education.

6.1.5 Challenge 4: Novel tasks for the TA

Similarly, the tasks envisaged for the drone were close to the ones found in HDI, surveillance (i.e., during exams) and safety Obaid et al. (2020a). While not often studied as social task, the participants thought of a companion, and a safely agent that could be there to guide and monitor.

6.1.5.1 Opportunity

Here again there will be opportunities to develop further research to provide social cues in this kind of scenario in order to inform students when they are being filmed or to guide them in a safe, private and comfortable manner.

6.1.6 Challenge 5: Social drones for classroom orchestration

Our participants mainly focused on the higher education setup (which is not well explored in the research literature for social robots in education). Several scenarios were described but often the social drone was envisioned to be used in a tutorial/workshop kind of setting rather than individual learning or classroom lecturing. This is an interesting aspect as the classroom orchestration and the teacher cognitive load managing group work needs to be taken into account when introducing technologies for the classroom Shahmoradi et al. (2019).

6.1.6.1 Opportunity

Future works could investigate how to leverage the social drone's high speed and mobility to use it for classroom orchestration (e.g., classroom monitoring, instructional scripting). An example of this orchestration is the management and monitoring of students' attention and well-being in the classroom that was often mentioned by participants. In these scenarios, participants thought of the drone as part of an Internet of Things ecosystem in which it could sense and adjust environmental aspects such as the light and the temperature.

6.2 Limitations and future directions

The above articulation on the design implications can support several future directions, which are also limitations in the presented study. For example, our focus is to gauge for the learners envisioned drone in education; taking a user (learner) centered design approach. This approach is also supported by others in field, such as the work presented by Reich-Stiebert et al. (2019a) and Lee et al. (2012a). Thus, the recruitment of our participants was done mainly through university student channels, and hence some of the graduate students came with teaching experience. However, we didn't select or expect our participants to have professional teaching experience, which one could consider as a limitation. In this context, we did notice that participants who had some teaching experience were the ones who thought of some teacher-centric scenarios "*Enables teachers to reflect on their quality of teaching*" (P05). Thus, conducting a study with school children and/or school teachers may further suggest additional valuable insights into the design of social drones in education. Moreover, despite the advantages afforded by our running an online design workshop, developing design insights from a face-to-face focus group workshop may reveal further implications towards social drones in education, as it could allow physical prototyping, such as the work presented by Karjalainen et al. (2017).

In summary, social robots in education are complex and challenging as they involve many stakeholders (i.e., students, teachers, parents, and the educational environment), individual and group interactions, and timely responses. As seen in previous research, drones offer a great potential for social interactions in the various application areas (Obaid et al. (2020a)). Thus, investigating the potential uses of social drones in education allowed us to generate novel perspectives for the design of social drones. We hope this work lays the foundation for the design of novel drones targeting teachers, learners, and the classroom environment.

Data availability statement

The datasets generated and analyzed in this study will be available on request to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by the Koç University Committee on Human Research Committee (No 2020.251. IRB3.092). The patients/participants provided their written informed consent to participate in this study.

Author contributions

WJ, DC, AY, and MO: Conception and design of the activity, data collection, data coding, literature search, interpretation of the results and critical revision of the manuscript for important intellectual content; AY: Ethics application and participants recruitment; DC and AY: Graphical design during and after the workshops; WJ: Implementation of the data analysis. WJ and MO: Drafting of the manuscript.

Funding

This work is partially funded by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation. We would also like to thank the Australian Research Council for their partial funding contribution (Grant No DE210100858).

Acknowledgments

We would like to thank Mafalda Gamboa for proof-reading the manuscript. We also would like to thank all the participants of the study for their valuable time.

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

Ahmad, M. I., Mubin, O., and Orlando, J. (2016). "Understanding behaviours and roles for social and adaptive robots in education: teacher's perspective," in Proceedings of the Fourth International Conference on Human Agent Interaction, Biopolis, Singapore (New York, NY, USA: Association for Computing Machinery), 297–304. HAI '16. doi:10.1145/2974804.2974829

Amirova, A., Rakhymbayeva, N., Yadollahi, E., Sandygulova, A., and Johal, W. (2021). 10 years of human-nao interaction research: a scoping review. *Front. Robot. AI* 8, 744526. doi:10.3389/frobt.2021.744526

Aydin, B. (2019). Public acceptance of drones: knowledge, attitudes, and practice. *Technol. Soc.* 59, 101180. doi:10.1016/j.techsoc.2019.101180

Barendregt, W., Ekström, S., Kiesewetter, S., Pareto, L., and Serholt, S. (2020). Demystifying robots in the co-design of a tutee robot with primary school children. *IxD&A.* 44, 109–128. doi:10.55612/s-5002-044-006

Baytas, M. A., Çay, D., Zhang, Y., Obaid, M., Yantaç, A. E., Fjeld, M., et al. (2019). Integrated Apparatus for Empirical Studies with Embodied Autonomous Social Drones. In 1st International Workshop on Human-Drone Interaction, Glasgow, UK 19, 1. doi:10.1145/3290605.3300480

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: a review. *Sci. Robot.* 3, eaat5954. doi:10.1126/scirobotics.aat5954

Bengtsson, M. (2016). How to plan and perform a qualitative study using content analysis. *NursingPlus Open* 2, 8–14. doi:10.1016/j.npls.2016.01.001

Cauchard, J. R., E, J. L., Zhai, K. Y., and Landay, J. A. (2015). "Drone & me: an exploration into natural human-drone interaction," in Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Association for Computing Machinery), 361–365. UbiComp. doi:10.1145/2750858.2805823

Cauchard, J. R., Zhai, K. Y., Spadafora, M., and Landay, J. A. (2016). "Emotion encoding in human-drone interaction," in 2016 11th ACM/IEEE international conference on human-robot Interaction, Christchurch, New Zealand, 263. doi:10.1109/HRI.2016.7451761

Chen, Y.-A., Wu, T.-Y., Chang, T., Liu, J. Y., Hsieh, Y.-C., Hsu, L. Y., et al. (2018). "ARPilot: Designing and investigating AR shooting interfaces on mobile devices for drone videography," in Proceedings of the 20th international conference on human-computer interaction with mobile devices and Services (Barcelona, Spain: Barcelona Association for Computing Machinery), 1. MobileHCI. doi:10.1145/3229434.3229475

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics Aut. Syst.* 42, 177–190. doi:10.1016/s0921-8890(02)00374-3

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382. Available at: https://www.statsmodels.org/dev/generated/statsmodels.stats.inter.rater.fleiss.kappa.html. doi:10.1037/h0031619

Hegel, F., Muhl, C., Wrede, B., Hielscher-Fastabend, M., and Sagerer, G. (2009). "Understanding social robots," in 2009 Second International Conferences on Advances in Computer-Human Interactions (IEEE), 169–174.

Herdel, V., Kuzminykh, A., Hildebrandt, A., and Cauchard, J. R. (2021). "Drone in love: emotional perception of facial expressions on flying robots," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.

Johnson, W. L., Rickel, J. W., Lester, J. C., et al. (2000). Animated pedagogical agents: face-to-face interaction in interactive learning environments. *Int. J. Artif. Intell. Educ.* 11, 47–78.

Johal, W. (2020). Research trends in social robots for learning. *Curr. Robot. Rep.* 1, 75–83. doi:10.1007/s43154-020-00008-3

Karjalainen, K. D., Romell, A. E. S., Ratsamee, P., Yantac, A. E., Fjeld, M., and Obaid, M. (2017). Social drone companion for the home environment: a user-centric exploration," in Proceedings of the 5th International Conference on Human Agent Interaction, Bielefeld, Germany, 17. Association for Computing Machinery, 89–96. doi:10.1145/3125739.3125774

Kennedy, J., Baxter, P., and Belpaeme, T. (2015). Comparing robot embodiments in a guided discovery learning interaction with children. *Int. J. Soc. Robot.* 7, 293–308. doi:10.1007/s12369-014-0277-4

Kim, H., and Landay, J. A. (2018). Aeroquake: Drone augmented dance." in Proceedings of the 2018 Designing Interactive Systems Conference, 18. Hong Kong: China: Association for Computing Machinery, 691–701. doi:10.1145/3196709.3196798

Knierim, P., Maurer, S., Wolf, K., and Funk, M. (2018). "Quadcopter-projected *in-situ* navigation cues for improved location awareness," in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada: Association for Computing Machinery), Montreal QC, 1. CHI. doi:10.1145/3173574.3174007

Köse, H., Uluer, P., Akalın, N., Yorgancı, R., Özkul, A., Ince, G., et al. (2015). The effect of embodiment in sign language tutoring with assistive humanoid robots. *Int. J. Soc. Robot.* 7, 537–548. doi:10.1007/s12369-015-0311-1

Lee, H. R., Sung, J., Šabanović, S., and Han, J. (2012a). "Cultural design of domestic robots: A study of user expectations in Korea and the United States," in 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, 803. doi:10.1109/ROMAN.2012.6343850

Lee, H. R., Sung, J., Sabanovic, S., and Han, J. (2012b). "Cultural design of domestic robots: A study of user expectations in korea and the United States," in IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication. Paris, France, 803-808. doi:10.1109/ROMAN.2012.6343850

Leong, T. W., and Johnston, B. (2016). "Co-Design and robots: a case study of a robot dog for aging people," in *Social robotics*. Editors A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He (Cham: Springer International Publishing), 702.

MagicLab, M. T. (2016). 24 Drone flight. last accessed on 05/10/2020

Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., and Dong, J.-J. (2013). A review of the applicability of robots in education. *J. Technol. Educ. Learn.* 1, 15. doi:10.2316/Journal.209.2013.1.209-0015

Mueller, F. F., and Muirhead, M. (2015). "Jogging with a quadcopter," in Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (Seoul, Republic of Korea: Association for Computing Machinery), 2023. CHI. doi:10.1145/2702123.2702472

Mumm, J., and Mutlu, B. (2011). "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in Proceedings of the 6th international conference on Human-robot interaction, Lausanne, Switzerland, 331–338.

Ng, W. S., and Sharlin, E. (2011). "Collocated interaction with flying robots," in *2011 ro-man*, 143. doi:10.1109/ROMAN.2011.6005280

Obaid, M., Aylett, R., Barendregt, W., Basedow, C., Corrigan, L. J., Hall, L., et al. (2018a). Endowing a robotic tutor with empathic qualities: design and pilot evaluation. *Int. J. Hum. Robot.* 15, 1850025. doi:10.1142/S0219843618500251

Obaid, M., Barendregt, W., Alves-Oliveira, P., Paiva, A., and Fjeld, M. (2015a). "Designing robotic teaching assistants: interaction design students' and children's views," in *Social RoboticsLecture notes in computer science*. Editors A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi (Springer International Publishing). 502–511. doi:10.1007/978-3-319-25554-5_50

Obaid, M., Baykal, G. E., Yantaç, A. E., and Barendregt, W. (2018b). Developing a prototyping method for involving children in the design of classroom robots. *Int. J. Soc. Robot.* 10, 279–291. doi:10.1007/s12369-017-0450-7

Obaid, M., Johal, W., and Mubin, O. (2020a). "Domestic drones: Context of use in research literature," in Proceedings of the 8th International Conference on Human-Agent Interaction (HAI '20), Sydney, Australia (Association for Computing Machinery). HAI '20. doi:10.1145/3406499.3415076

Obaid, M., Kistler, F., Häring, M., Bühling, R., and André, E. (2014). A framework for user-defined body gestures to control a humanoid robot. *Int. J. Soc. Robot.* 6, 383–396. doi:10.1007/s12369-014-0233-3

Obaid, M., Kistler, F., Kasparavičiūtundefined, G., Yantaç, A. E., and Fjeld, M. (2016a). "How would you gesture navigate a drone? a user-centered approach to control a drone," in Proceedings of the 20th International Academic Mindtrek Conference, (Tampere, Finland: Association for Computing Machinery), AcademicMindtrek 113–121. doi:10.1145/2994310.2994348

Obaid, M., Mubin, O., Basedow, C. A., Ünlüer, A. A., Bergström, M. J., and Fjeld, M. (2015b). "A drone agent to support a clean environment," in Proceedings of the 3rd International Conference on Human-Agent Interaction (Daegu, Kyungpook, Republic of Korea, Daegu, Kyungpook, Republic of Korea) (Association for Computing Machinery). HAI '15, 55–61. doi:10.1145/2814940.2814947

Obaid, M., Mubin, O., Brown, S. A., Yantac, A. E., Otsuki, M., and Kuzuoka, H. (2020b). "DroEye: introducing a social eye prototype for drones," in Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. (Cambridge, United Kingdom: Association for Computing Machinery), 378–380. doi:10.1145/3371382.3378313

Obaid, M., Yantaç, A. E., Barendregt, W., Kırlangıç, G., and Göksun, T. (2016b). "Robo2box: a toolkit to elicit children's design requirements for classroom robots," in *Social robotics*. Editors A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He (Kansas City, Cham: Springer International Publishing), 600–610.

Reich-Stiebert, N., Eyssel, F., and Hohnemann, C. (2019a). Exploring university students' preferences for educational robot design by means of a user-centered design approach. *Int. J. Soc. Robot.* 12, 227–237. doi:10.1007/s12369-019-00554-7

Reich-Stiebert, N., Eyssel, F., and Hohnemann, C. (2019b). Involve the user! changing attitudes toward robots by user participation in a robot prototyping process. *Comput. Hum. Behav.* 91, 290–296. doi:10.1016/j.chb.2018.09.041

Rohrbach, B. (1969). *Methode 635, eine neue technik zum lösen von problemen*. German sales magazine "Absatzwirtschaft".

Romanowski, A., Mayer, S., Lischke, L., Grudzieundefined, K., Jaworski, T., Perenc, I., et al. (2017). "Towards supporting remote cheering during running races with drone technology," in Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems, Denver, Colorado (New York, NY, USA: Association for Computing Machinery), (CHI EA), 2867–2874. doi:10.1145/3027063.3053218

Rubens, C., Braley, S., Torpegaard, J., Lind, N., Vertegaal, R., and Merritt, T. (2020). "Flying LEGO bricks: observations of children constructing and playing with programmable matter," in Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction (Sydney NSW, Australia: Association for Computing Machinery), (TEI), 193–205. doi:10.1145/3374920.3374948

Saerbeck, M., Schut, T., Bartneck, C., and Janse, M. D. (2010). "Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia (New York, NY, USA: Association for Computing Machinery), (CHI '10), 1613–1622. doi:10.1145/1753326.1753567

Schäffer, B., Pieren, R., Heutschi, K., Wunderli, J. M., and Becker, S. (2021). Drone noise emission characteristics and noise effects on humans—a systematic review. *Int. J. Environ. Res. Public Health* 18, 5940. doi:10.3390/ijerph18115940

Scheible, J., and Funk, M. (2016). "*In-situ*-displaydrone: facilitating co-located interactive experiences via a flying screen," in Proceedings of the 5th ACM International Symposium on Pervasive Displays, Oulu, Finland (New York, NY, USA: ACM PerDis), 251–252. doi:10.1145/2914920.2940334

Schneegass, S., Alt, F., Scheible, J., and Schmidt, A. (2014). "Midair displays: concept and first experiences with free-floating pervasive displays," in Proceedings of The International Symposium on Pervasive Displays, Copenhagen, Denmark (New York, NY, USA: PerDis), 2731. 27–27. doi:10.1145/2611009.2611013

Serholt, S., Basedow, C. A., Barendregt, W., and Obaid, M. (2014). "Comparing a humanoid tutor to a human tutor delivering an instructional task to children," in Proceedings of the 2014 IEEE-RAS International Conference on Humanoid Robots, Madrid, Spain, 1134–1141.

Shahmoradi, S., Olsen, J. K., Haklev, S., Johal, W., Norman, U., Nasir, J., et al. (2019). "Orchestration of robotic activities in classrooms: challenges and opportunities," in European Conference on Technology Enhanced Learning, Delft, (Netherlands (Cham): Springer), 640–644.

Szafir, D., Mutlu, B., and Fong, T. (2015). "Communicating directionality in flying robots," in Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, Portland, Oregon (New York, NY, USA: Association for Computing Machinery), 19–26. HRI '15. doi:10.1145/2696454.2696475

Šabanovic, S., Chang, W. L., Bennett, C. C., Piatt, J. A., and Hakken, D. (2015). "A robot of my own: participatory design of socially assistive robots for independently living older adults diagnosed with depression," in *HCI*.

Tezza, D., and Andujar, M. (2019). The state-of-the-art of human–drone interaction: a survey. *IEEE Access* 7, 167438–167454. doi:10.1109/access.2019.2953900

Treurniet, T., Bai, L., à Campo, S., Wang, X., Hu, J., and Barakova, E. (2019). "Drones with eyes: expressive human-drone interaction," in 1st International Workshop on Human-Drone Interaction (Ecole Nationale de l'Aviation Civile [ENAC]), Glasgow, UK.

Walker, M., Hedayati, H., Lee, J., and Szafir, D. (2018). "Communicating robot motion intent with augmented reality," in Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (Chicago, IL, USA: Association for Computing Machinery), (HRI '18), 316–324. doi:10.1145/3171221.3171253

Watkins, S., Burry, J., Mohamed, A., Marino, M., Prudden, S., Fisher, A., et al. (2020). Ten questions concerning the use of drones in urban environments. *Build. Environ.* 167, 106458. doi:10.1016/j.buildenv.2019.106458

Yanco, H. A., and Drury, J. (2004). "Classifying human-robot interaction: an updated taxonomy," in 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), The Hague, Netherlands, October, 3, 2841–2846.

Yeh, A., Ratsamee, P., Kiyokawa, K., Uranishi, Y., Mashita, T., Takemura, H., et al. (2017). "Exploring proxemics for human-drone interaction," in Proceedings of the 5th international conference on human agent interaction (Bielefeld, Germany: Association for Computing Machinery), 17, 81–88. doi:10.1145/3125739.3125773

Yılmaz, R., and Kılıç-Çakmak, E. (2012). Educational interface agents as social models to influence learner achievement, attitude and retention of learning. *Comput. Educ.* 59, 828–838. doi:10.1016/j.compedu.2012.03.020

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership