Second Language Tutoring using Social Robots

Project No. 688014

L2TOR

Second Language Tutoring using Social Robots

Grant Agreement Type: Collaborative Project
Grant Agreement Number: 688014

D8.3 Open Publication

Due Date: **31/12/2018**
Submission Date: **14/01/2019**

Start date of project: **01/01/2016** Duration: **36 months**

Organisation name of lead contractor for this deliverable: **Plymouth University**

Responsible Person: **Tony Belpaeme** Revision: **1.0**

Project co-funded by the European Commission within the H2020 Framework Programme		
Dissemination Level		
PU	Public	**PU**
PP	Restricted to other programme participants (including the Commission Service)	
RE	Restricted to a group specified by the consortium (including the Commission Service)	
CO	Confidential, only for members of the consortium (including the Commission Service)	

Contents

Executive Summary

This deliverable consists of all publications resulting from the L2TOR project over the past three years. All references in this deliverable are categorised by year of publication. These reference lists are followed by the full publications, which are listed in the same order as the references.

1 2018 Publications

- Rianne van den Berghe, Josje Verhagen, Ora Oudgenoeg-Paz, Sanne van der Ven, and Paul Leseman (2018). Social robots for language learning: A review. *Review of Educational Research* - In Press *

- Bahar Irfan, James Kennedy, Séverin Lemaignan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme (2018, March). Social Psychology and Human-Robot Interaction: An Uneasy Marriage. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (pp. 13-20).* ACM, Chicago, IL, USA.

- Bahar Irfan, Natalia Lyubova, Michael G. Ortiz, and Tony Belpaeme (2018). Multi-modal Open-Set Person Identification in HRI. *Social Robots in the Wild workshop at HRI2018.*

- Emmy Rintjema, Rianne van den Berghe, Anne Kessels, Jan de Wit, and Paul Vogt. (2018). A Robot Teaching Young Children a Second Language: The Effect of Multiple Interactions on Engagement and Performance. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (pp. 219-220).* ACM, Chicago, IL, USA.

- Rianne van den Berghe, Sanne van der Ven, Josje Verhagen, Ora Oudgenoeg-Paz, Fotios Papadopoulos, and Paul Leseman. (2018). Investigating the Effects of a Robot Peer on L2 Word Learning. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (pp. 267-268).* ACM, Chicago, IL, USA.

- Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt (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 (HRI '18).* ACM, Chicago, IL, USA, 50-58.

- Tony Belpaeme, Paul Vogt, Rianne van den Berghe, Kirsten Bergmann, Tilbe Göksun, Mirjam de Haas, Junko Kanero, James Kennedy, Aylin C. Küntay, Ora Oudgenoeg-Paz, Fotios Papadopoulos, Thorsten Schodde, Josje Verhagen, Christopher D. Wallbridge, Bram Willemsen, Jan de Wit, Vasfiye Geçkin, Laura Hoffmann, Stefan Kopp, Emiel Krahmer, Ezgi Mamus, Jean-Marc Montanier, Cansu Oranç, and Amit Kumar Pandey (2018) Guidelines for Designing Social Robots as Second Language Tutors. *International Journal of Social Robotics.* Springer Verlag, Germany, 10/3.

*No pdf available - in press

- Junko Kanero, Vasfiye Geçkin, Cansu Oranç, Ezgi Mamus, Aylin C. Küntay, and Tilbe Göksun (2018), Social Robots for Early Language Learning: Current Evidence and Future Directions. *Child Development Perspectives.* Wiley-Blackwell, USA, 12/3.

- Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka (2018). Social robots for education: A review. *Science Robotics.* AAAS, Washington DC, USA, 3/21.

- Christopher D. Wallbridge, Rianne van den Berghe, Daniel Hernández Garcia, Junko Kanero, Séverin Lemaignan, Charlotte Edmunds, and Tony Belpaeme (2018). Using a Robot Peer to Encourage the Production of Spatial Concepts in a Second Language. In *Proceedings of the 6th International Conference on Human-Agent Interaction – HAI '18.* ACM, New York, NY, USA, 54-60

- Thorsten Schodde and Stefan Kopp (2018). Adaptive Robot Second Language Tutoring for Children. Companion of *the 2018 ACM/IEEE International Conference on Human-Robot Interaction – HRI '18.* ACM, New York, NY, USA, 317-318. *

- Paul Vogt, Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, and Emiel Krahmer (2018). Iconic gestures improve second language learning from a social robot. Presented at *the ISGS: International Society for Gesture Studies,* Cape Town, South Africa. *

- Pieter Wolfert, Mirjam de Haas, Paul Vogt, and Pim Haselager (2018). Measuring Engagement Online in CRI. In *Proceedings of The Near Future of Children's Robotics, IDC Workshop 2018,* Trondheim, Norway.

- Jan de Wit, Bram Willemsen, Mirjam de Haas, Pieter Wolfert, Paul Vogt, and Emiel Krahmer (2018). Playful exploration of a robot's gesture production and recognition abilities. Poster presented at *Workshop Gesture & Technology 2018,* Warwick, United Kingdom.

- Laura Hoffmann, Sonja Stange, Thorsten Schodde, and Stefan Kopp. (2018). How Transparency can Foster Second Language Learning for (Some) Children. *When Robots Think – Interdisciplinary Views on Intelligent Automation Symposium* 14. – 17.11.18, Münster, Germany. *

- Séverin Lemaignan, Charlotte E. R. Edmunds, Emmanuel Senft, and Tony Belpaeme (2018). The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PLOS ONE* 13(10).

*No pdf available

- Séverin Lemaignan, Yoan Sallami, Christopher Wallbridge, Aurélie Clodic, Tony Belpaeme, and Rachid Alami (2018). UNDERWORLDS: Cascading Situation Assessment for Robots. In *Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems,* Madrid, Spain.

- Anna-Lisa Vollmer, Robin Read, Dries Trippa, and Tony Belpaeme (2018). Children conform, adult resist: A robot group induced peer pressure on normative social conformity. *Science Robotics,* 3/21.

- Christopher D. Wallbridge, Séverin Lemaignan, Emmanuel Senft, Charlotte Edmunds and Tony Belpaeme, 2018, March. Spatial Referring Expressions in Child-Robot Interaction: Let's Be Ambiguous! *4th Workshop on Robots for Learning (R4L) – Inclusive Learning @HRI 2018.* Chicago, IL, USA.

- Rianne van den Berghe, Paul Vogt, Junko Kanero, Hanneke Leeuwestein, and Ora Oudgenoeg-Paz. (2018, August). Social Robots for Language Learning. Symposium paper presented at *the EARLI SIG 5 Conference 2018 – Learning and Development in Early Childhood,* Berlin, Germany.*

- Paul Vogt, Bram Willemsen, Jan de Wit, Mirjam de Haas, and Emiel Krahmer. (2018, August). Personalized and multimodal interactions for second language tutoring using a social robot. Social Robots for Language Learning. Symposium paper presented at *the EARLI SIG 5 Conference 2018 – Learning and Development in Early Childhood,* Berlin, Germany.*

- Hanneke Leeuwestein, Marie Barking, Hande Sodaci, Rian Aarts, Jan de Wit, Ora Oudgenoeg-Paz, Josje Verhagen, and Paul Vogt (2018, August). Bilingual robots teaching L2 vocabulary to immigrant children. Symposium paper presented at *the EARLI SIG 5 Conference 2018 – Learning and Development in Early Childhood,* Berlin, Germany. *

- Rianne van den Berghe, Sanne van der Ven, Josje Verhagen, Ora Oudgenoeg-Paz, Fotios Papadopoulos, and Paul Leseman (2018, August). The effect of a robot peer on second language vocabulary learning gains. Symposium paper presented at *the EARLI SIG 5 Conference 2018 – Learning and Development in Early Childhood,* Berlin, Germany. *

- Junko Kanero, Özlem Ece Demir-Lira, Sümeyye Koşkulu, Cansu Oranç, Idil Franko, Aylin C. Küntay, and Tilbe Göksun (2018, August). How do robot gestures help second language learning? Presented at *the EARLI SIG 5 Conference 2018 – Learning and Development in Early Childhood,* Berlin, Germany.

*No pdf available

- Sümeyye Koşkulu, Cansu Oranç, Junko Kanero, Tilbe Göksun, and Aylin C. Küntay (2018). Robotların Jestler ve Geri Bildirim İle Okul Öncesi Dönemde İngilizce Eğitimine Katkıları. In *Proceedings of Ulusal Psikoloji Kongresi,* Ankara, Turkey.

- Ö. Ece Demir-Lira, Junko Kanero, Cansu Oranç, Sümeyye Koşkulu, Idil Franko, Zeynep Adıgüzel, Orhun Uluşahin, Tilbe Göksun, and Aylin C. Küntay (2018). Using gestures in L2 vocabulary teaching: Human or robot tutors? Poster presented at *The 43rd Annual Boston University Conference on Language Development (BUCLD),* Boston, Massachusetts, USA.

- Junko Kanero, Idil Franko, Cansu Oranç, Orhun Uluşahin, Sümeyye Koşkulu, Zeynep Adıgüzel, Aylin C. Küntay, and Tilbe Göksun (2018). Who can benefit from robots? Effects of individual differences in robot-assisted language learning. In *Proceedings of the 8th Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics,* Tokyo, Japan.

- Ö. Ece Demir-Lira, Cansu Oranç, Junko Kanero, Sümeyye Koşkulu, Idil Franko, Zeynep Adıgüzel, Tilbe Göksun, and Aylin C. Küntay (2018). Sosyal robotların jest kullanımının çocuklarda ikinci dil öğrenimine. Poster presented at *Türkiye Robotbilim Konferansı (ToRK),* Istanbul, Turkey.

Social Psychology and Human-Robot Interaction: An Uneasy Marriage

Bahar Irfan
CRNS, Plymouth University, UK
bahar.irfan@plymouth.ac.uk

James Kennedy
CRNS, Plymouth University, UK
james.kennedy@plymouth.ac.uk

Séverin Lemaignan
CRNS, Plymouth University, UK
severin.lemaignan@plymouth.ac.uk

Fotios Papadopoulos
CRNS, Plymouth University, UK
fotios.papadopoulos@plymouth.ac.uk

Emmanuel Senft
CRNS, Plymouth University, UK
emmanuel.senft@plymouth.ac.uk

Tony Belpaeme
ID Lab, Ghent University, Belgium
CRNS, Plymouth University, UK
tony.belpaeme@plymouth.ac.uk

ABSTRACT

The field of Human-Robot Interaction (HRI) lies at the intersection of several disciplines, and is rightfully perceived as a prime interface between engineering and the social sciences. In particular, our field entertains close ties with social and cognitive psychology, and there are many HRI studies which build upon commonly accepted results from psychology to explore the novel relation between humans and machines. Key to this endeavour is the trust we, as a field, put in the methodologies and results from psychology, and it is exactly this trust that is now being questioned across psychology and, by extension, should be questioned in HRI.

The starting point of this paper are a number of failed attempts by the authors to replicate old and established results on social facilitation, which leads us to discuss our arguable over-reliance and over-acceptance of methods and results from psychology. We highlight the recent "replication crisis" in psychology, which directly impacts the HRI community and argue that our field should not shy away from developing its own reference tasks.

KEYWORDS

Social psychology; social robotics; replication crisis; Human-Robot Interaction; research methodology

ACM Reference Format:
Bahar Irfan, James Kennedy, Séverin Lemaignan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2018. Social Psychology and Human-Robot Interaction: An Uneasy Marriage. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3173386.3173389

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03…$15.00
https://doi.org/10.1145/3173386.3173389

1 THE REPLICATION CRISIS IN PSYCHOLOGY AND WHAT IT MEANS FOR HRI

The field of Human-Robot Interaction (HRI), and in particular, the field of *social* HRI benefits from a wide range of scientific input [4, 5]. As a community, we recognise that the technical fields of engineering, control theory and computer science do not provide necessary tools for the scientific investigation of the 'human' and 'interaction' parts of HRI. For this reason, we take inspiration and ground much of our research in established results from the social sciences – primarily social psychology, cognitive psychology, and sociology. As scholars in HRI we find ourselves at the intersection of these many fields, and aim to offer insights to programmers and engineers, as well as psychologists. In this sense, our field embodies the basic idea of cognitive sciences: building bridges across disciplines to gain new insights on complex scientific challenges.

That said, the demographics of the academics working in HRI are skewed towards engineering backgrounds (Table 1); one often becomes a researcher in HRI by first building robots and then looking at how the machines might interact with humans. While some of us do have training in psychology, many do not. This is not an issue per se: as trained scientists and engineers, we can read and interpret the social science literature, and reproduce tasks, protocols, and –perhaps– results.

However, the recent replication crisis in psychology now casts doubt on that premise. Aarts et al. [1], in their seminal study, found that upon attempting to replicate 100 psychology studies, only 39% of the replication studies could subjectively be rated to have replicated the original result. As the results of two thirds of 100 studies could not be properly replicated, whatever the reasons might be (from publication bias, to sociological changes in the population,

Table 1: Academic fields of accepted authors at HRI17, as judged by their affiliation or, if advertised on their personal website, training, $n = 193$ (a single author can be affiliated with multiple fields).

Field	*Eng.*	*Psy.*	*Cog. Sci.*	*Interaction Design*	*Other*
N	145	24	6	17	13
%	70.73	11.71	2.93	8.29	6.34

to small effect sizes), it calls for exerting caution whenever we build upon supposedly established results.

Further research has shown that many scientific studies are difficult or impossible to replicate upon subsequent investigation. According to a 2016 poll of almost 1,600 scientists reported in the journal Nature, over 70% had failed to reproduce at least one other scientist's experiment. More than half had failed to reproduce one of their own experiments [3]. This is problematic for the field of Human-Robot Interaction, as much of what we do either uses research methods similar to those used in other disciplines (and psychology in particular), or relies directly on insights and results handed down from other disciplines.

Because many of us are *consumers* of the psychology literature rather than *producers* or active contributors to the psychology community, we often do not only have insufficient training to correctly interpret psychological studies, but also tend to be less critical and often do not question findings the way we would in our own community. This effect is reinforced by the perceived maturity of different academic fields. Fields such as social or cognitive psychology are very mature, compared to the relative immaturity of Human-Robot Interaction, and studies and insights from psychology are now core material in textbooks, giving the studies and their results further credence.

While experienced researchers in HRI might already be aware of these issues, the influx of new talent requires our field to be vigilant of uncritical reliance on questionable methods and results. To illustrate our point, we present our experience in which we were unsuccessful at reproducing the *social facilitation* effect. Social facilitation, also known as the audience effect, is a supposedly well-established effect where the mere presence of a (silent, passive) external agent influences one's behaviour, often measured through performance on a task. The direction in which the effect works is not specified: depending on the task and the context, performance can be positively or negatively impacted. A large body of literature from psychology reports this effect, and social facilitation has been studied in robotics as well in various forms.

2 A CASE IN POINT: SOCIAL FACILITATION

2.1 Context: Studying the Mere Presence Effect in Social Facilitation

Background and Related Theories. In 1898, Triplett [37] observed that cyclists pedal faster in the presence of rivals than when they are alone. He later studied this effect on children by using a fishing reel that they needed to turn as quickly as possible and found the same effect, although a later analysis of his work by Stroebe [34] showed that there was no significant difference in either of his findings. This effect has later been termed as 'social facilitation' by Allport [2] to describe the increase in response due to the presence of others who are performing the same task. Later the term social facilitation was expanded to cover two types of conditions: 'co-action effects' like Triplett's examples, and 'audience effects', in which only the mere presence of an observer affects the performance of a person performing the task. In order to explain the audience effects, Zajonc [42] proposed the *drive theory*, which states that the audience enhances the exhibition of dominant responses in a person. In the case of a well-mastered task ('simple task'), the performance is facilitated, whereas, for the tasks that are new or require learning ('complex tasks'), the performance is inhibited.

Factors. A meta-analysis by Bond and Titus [8] compared 202 published and 39 unpublished studies on social facilitation. They provide a list of 13 factors that might impact social facilitation (like the participants' age, the number of observers, the role of the observers, the familiarity of the observers, etc.). The meta-study shows that the performance speed (*quantity*) is increased for the simple tasks and the performance accuracy (*quality*) is decreased for the complex tasks. The performance quantity is measured by the latency to respond, time it takes to complete a task and the number of responses per unit time. The performance quality is measured by the number of errors. The analysis also showed that the visibility (presence in the same room as the subject) of the observers has a slightly larger effect than the non-visibility (e.g., one-way mirror [11, 14], use of a video camera [16, 36], a desktop image on a computer screen [15]), although the difference was not statistically significant.

On the other hand, Guerin [17] argues in his review that the majority of studies on social facilitation had observers watching the subject perform a task. These could be confederates, but often they are just the experimenter watching a subject, as they were not seen being busy with other tasks. He also draws attention to ceiling and floor effects of the tasks, and advises that the task should be sufficiently hard so that a reasonable comparison can be made between subjects and conditions.

Tasks. Following Zajonc [42], the literature on social facilitation distinguishes between 'simple tasks' and 'complex tasks'. Examples of simple tasks include cancelling specific letters in a text or multiplication; examples of complex tasks include concept formation, anagrams, digit span, and pursuit rotor tasks (a motor task in which the subject has to track a rotating target using a computer mouse). Tasks such as letter copying and paired associates can be either simple or complex depending on the task structure. McCaffrey et al. [23] also presented significance levels of each of these tasks in the literature. They show that visual perception and construction tasks such as letter or word copying [15, 18, 36] and motor tasks such as physical activities [35] are good tasks in terms of significance as simple tasks. Memory or learning tasks such as paired associates [10, 16, 17] and visuomotor tasks as in the rotary pursuit task [22, 25] have higher significance for social facilitation as complex tasks.

Cheating as a reinforcing factor. Self-presentation theory [7] also suggests conformity to normative behaviours to gain approval of another person. For example, in the case of an embarrassing situation such as cheating, this should prevent the subject from engaging in the cheating behaviour due to social pressure. There might be several factors that affect cheating behaviour, such as the importance of the task, the risk of being caught, the probability of success [39], the belief in free-will [40], the knowledge of peer performance [19], the potential gain of money or grades, the penalty for cheating [26] or conformity to cheating behaviour in peers [13]. In the study by Vohs and Schooler [40], the task consisted of a computer-based mental arithmetic test. The participants were told that there was a "glitch" in the program which shows the correct

answer to the problem, but they could close the answer window by pressing a key after the problem appeared. They were also told that the experimenter would not know whether they pressed the bar, but they should try to solve the problems without looking at the answer. The results revealed that those who were given an essay prior to the test that stated the lack of free-will cheated more frequently than others.

Social Facilitation In Robotics. The audience effect has been studied in HRI by Schermerhorn et al. [32] and Riether et al. [27]. Schermerhorn et al. [32] compared the effect of the robot's presence during easy and difficult arithmetic tasks with alone and robot-presence conditions. A significant two-way interaction between gender and robot was found, because the subjects performed worse during the difficult task when the robot was present. Overall, a marginally significant effect of robot presence was found. Riether et al. [27] on the other hand, compared alone, human-presence, and anthropomorphic robot-presence conditions with four different tasks with easy and complex conditions: anagram solving, numerical distance, finger tapping and a motor reaction task. They observed that in the anagram solving, numerical distance and finger tapping tasks, there were significantly larger performance scores than the alone group for both the robot and human conditions, but there was no significant difference between the robot and the human observer conditions. Authors concluded that this finding suggests that people regard robots as social beings. After the experiment, the subjects were asked to complete a questionnaire in which they gave higher observation impression scores for the robot condition than the human observer, perhaps due to the fact that they thought someone else was watching through the eyes of the robot or due to novelty effects leading to distraction.

Following the findings from social facilitation literature, we decided to explore the mere presence of two robotic platforms (the Softbank Robotics NAO and Pepper robots) through a social facilitation task. We anticipated that there would be a difference between the two platforms due to their size and appearance. While the studies aimed to compare the social facilitation of two different robots, it was important to establish two baselines first: one with no observers, and one with the social facilitation elicited by the presence of a human observer. This would essentially be a first step in validating our methodology and would also serve as replication of the finding from psychology. Assuming the replication study was successful, we would have continued the experiments with a robot as observer and would have compared these results to the earlier obtained baselines.

We ran two separate studies, with a total of three different tasks. Because no effect could be found between the alone condition and the human condition in any of our tasks, we did not actually pursue the studies with robots.

2.2 Social Facilitation: First Attempt

The first study was run between-subjects with two conditions: an alone condition and a human-presence condition. Participants were recruited on a university campus and taken to a room in the campus library for the experiment. The experimenter would take the participant to the room and tell them to follow instructions on the tablet, then the experimenter would leave. In the human observer condition, a second experimenter would already be sitting in the room and would remain there for the duration of the experiment (as per Figure 1).

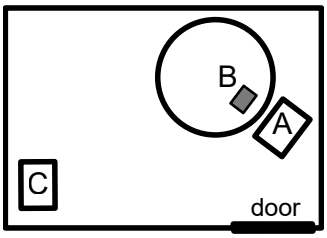


Figure 1: Layout of the room. The participant (*A*) is sitting at a table, with their back to the door. The tasks are performed on a tablet (*B*). When present, the social agent (human observer) is placed at *C*.

Tasks. The literature distinguishes between the effects of mere presence on *simple tasks* and *complex tasks*. We sought to elicit differences in both of these task types. Each participant therefore performed two tasks, followed by a brief questionnaire. Both the tasks and the questionnaire were administered on the tablet. The first task was designed to be a repetitive visuomotor task (the 'shape matching' task); the second one required recollection and comprehension of spoken information (the 'story' task). As such, we examined the effect of social facilitation on both low- and high-cognitive tasks.

The 'shape matching' task is a game where the participants are asked to match a coloured target shape with another one, of the same shape, but of a different colour (Figure 2). The target shape as well as the eight possible responses are random combinations from the sets {red, yellow, green, purple, blue, white} and {square, cross, star, circle}. After the participant touches a shape to select it as an answer, a new random set is shown on screen. This is repeated 200 times. By using the same random seed for all participants, the stimuli sequence was kept identical for all participants.

The task can be repeated for up to 200 rounds of random shapes. After 75 rounds, a button labelled "Give up" appears on screen, giving the participant the option to skip to the second task. The wording of the label was intentionally chosen instead of a more neutral "Stop" or "Continue to next task" to elicit a stronger social response ("Giving up" being more socially costly than simply "continuing to the next task"), thereby increasing the contrast between conditions (self-presentation effect). During this first task, we recorded three metrics: the reaction time for each round, the number of correct and incorrect responses, and the total number of rounds completed. We also asked the participants to give an estimate of how many rounds they thought they had completed, between 0 and 300.

The second task ('story' task) involves listening to a short pre-recorded text (1min 56sec) and answering eight questions about this text. The text[1] details the history of a fictional country named "Brookland" and includes a range of facts: names of places ("[they] sailed to Port Danford"), dates ("Springland was settled in the year

[1]Recording and transcript available on-line, at https://github.com/severin-lemaignan/shapes-matching/tree/master/audio.

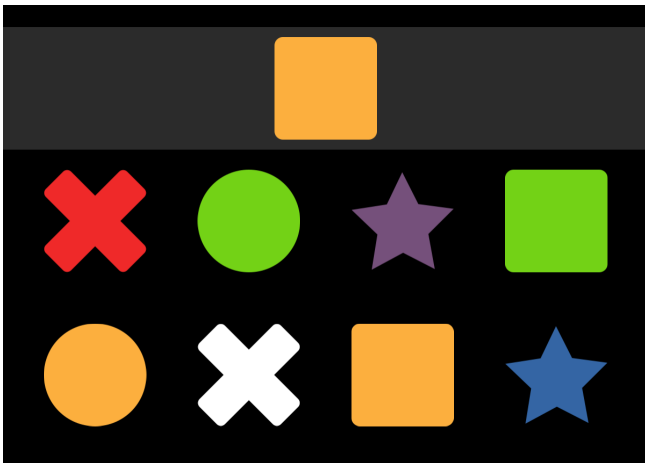


Figure 2: Screenshot of the *shape matching* task. Participants are instructed to tap on the picture matching the target's shape (seen at the top), but with a different colour. In this example, the participant has to tap the green square.

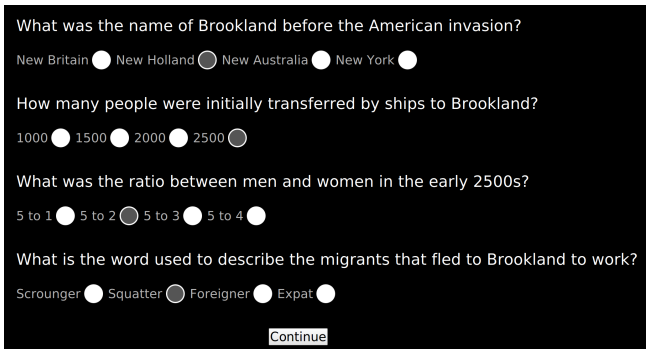


Figure 3: Screenshot of four of the eight questions administered immediately after listening to a short text about a fictional country.

2503"), terminology ("Settlers or 'squatters' began to move deeper into the territories"), and situations ("Women were outnumbered five to one"). The text was based upon the settling of Australia, but with key details and place names changed. This was so that the information would certainly be novel, without sounding implausible. The eight multiple-choice questions are asked immediately after the end of the text. Each question provides a choice of four answers (Figure 3). The score of each participant (number of correct answers) is the performance metric for this task.

Hypotheses. Based on the *drive* theory by Zajonc [42], our hypotheses were the following:

- H1 In the 'shape matching' task, the presence of a social agent would lead to **better performance**: fewer mistakes, faster reaction times.
- H2 In the presence of a social agent, the 'Give up' button would be used less frequently (or later in the game) due to the social pressure (self-presentation theory).
- H3 In the presence of a social agent, participants would report that they completed fewer rounds of shape matching than they actually did, due to social facilitation.
- H4 In the 'story' task, the presence of a social agent would lead to the impaired performance, i.e., participants would recall fewer facts.

Protocol & Data Collection. We recruited 45 participants after exclusion (25 for the alone condition and 20 for the human condition, 16 males, 29 females, balanced across conditions) on campus. The participants' age was M=20.4 (SD=2.5). We ensured that all participants who enrolled were not colour-blind (due to the necessity of seeing colour accurately for the shape matching task) and that they were native English speakers (to prevent comprehension issues due to language in the story task).

Participants were first given information sheets describing the experiment (simply entitled "Learning with a touchscreen", so as not to disclose the role of the mere presence of the observers). They then gave consent to participate, compliant with the university ethics committee rules. Participants were told in writing and verbally that whether or not they decided to withdraw early from the study, they would receive financial compensation of £5 (in the form of a voucher). We made this point explicit to make sure the participants knew that, even if they quit the shape matching game early (i.e., between rounds 75 and 200), they would still receive the full amount.

Results. We did not observe any significant difference between the two conditions concerning the time required to match 75 shapes, average reaction time, number of shape completed, ratio of correct matching, recall performance, or perceived observation (cf. Table 2).

This means that we did not observe any social facilitation effect, and none of the hypotheses are supported.

2.3 Second Attempt

Reflecting on the lack of effect observed in our first attempt, we designed a second experiment to address the possible failures of the first one.

Specifically, we chose (1) to have the human observer closer to the participant (aiming for greater human influence), (2) a stronger moral component (aiming for a greater influence of the human presence), (3) a more difficult task (stronger incentive for behavioural differences – i.e., cheating – between conditions), (4) financial reward dependent on performance (stronger, clearer incentive for behavioural differences between conditions) and finally, (5) regarding the methodology, we decided to move away from primarily using reaction times as metric, so as to avoid any natural performance limit.

Task. Based on these constraints, we designed a new task involving mental arithmetic. Participants were required to calculate the result of a set of non-trivial mental additions. The additions each had exactly three 2-digit numbers to sum, one carry (a digit that is transferred from one column of digits to another), and their results ranged from 100 to 200. Participants had 5 minutes to perform as many additions as possible. Each correct answer would earn them a small financial reward of £0.20 (Figure 4).

Table 2: Results for the shape matching task: time to match 75 shapes, average reaction time, number of shapes completed, ratio of perceived matching, recall performance, and perceived observation. No significance has been observed for any of the metrics (2-tailed independent 2-samples test with equal variance assumption).

Metric	Alone condition $M(SD)$	Human condition $M(SD)$	*p-value*	*t-value*
Time to 75 shapes (s)	117.7 (30.01)	110.63 (17.25)	.349	0.948
Average reaction time (s)	1.70 (0.47)	1.58 (0.26)	.305	1.037
Number of shapes completed	196 (11.5)	198 (7.8)	.522	−0.596
Ratio of correct responses	0.98 (0.02)	0.99 (0.01)	.082	−1.813
Recall performance	4.81 (1.27)	5.11 (1.49)	.473	−0.724
Perceived observation	2.76 (1.27)	2.55 (1.39)	.6	0.528

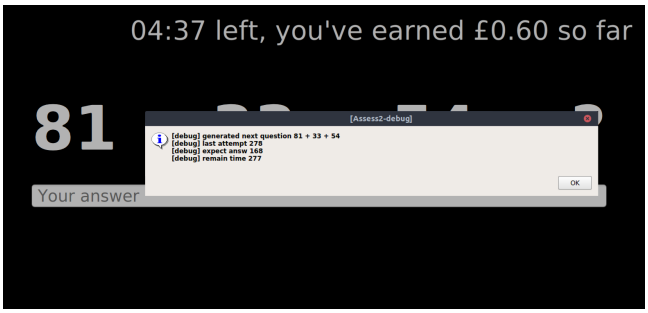


Figure 4: Screenshot of the *sums* task. Participants have 5 minutes to perform as many mental additions of three numbers (with a result between 100 and 200) as possible. Each correct answer earns £0.20. A pop-up dialogue with the answer is shown before each new addition. The participants are instructed that this pop-up is a bug and that they should ignore it and dismiss it.

Critically, following the design of Vohs and Schooler [40], a supposed "glitch" was showing a pop-up dialogue before each addition. This dialogue was designed to look like a spurious debug dialogue and contained the expected answer. The participants were explicitly shown by the experimenter that the correct answer was erroneously displayed in the dialogue. They were instructed to ignore the dialogue and to dismiss it. This 'bug' was explained to the participant as being caused by a new operating system on the laptops used for the test ("Our previous computers did not have this issue"). The bug made it practically easy for participants to cheat: by briefly glimpsing at the debug dialogue before dismissing it, they could immediately know the correct answer, and earn money faster.

The dialogue could be dismissed by pressing 'enter' on the keyboard. 'Enter' was also the key used to move to the next question. As such, a double-press would move to the next question and close the dialogue before it could be seen. Through this mechanism, it was possible to measure how long it took participants to close the dialogue, and infer whether they had cheated.

Hypotheses. The literature suggests that social presence during a complex cognitive task like this one should lead to worse performance [12, 41]. Accordingly, our hypotheses were the following:

- H1 In the presence of a social agent, participants will be more honest (i.e., they will look at the answer on the dialogue pop-up less).
- H2 In the presence of a social agent, participants will complete fewer correct questions.

Protocol & Data Collection. As outlined previously, while our plan was to run four conditions (alone, human presence, NAO presence, Pepper presence), we first ran the two baseline conditions: alone and with a human observer. 15 participants were recruited in the alone condition, 16 participants in the human condition.

The experimental setup was similar to Figure 1 with two differences: when present, the human observer was sitting at the table, facing the participant, and the tablets were replaced with laptops with a keyboard to facilitate the input of the answers. For each participant, we recorded how many additions were attempted, the total gain (i.e., the number of correct answers), and the time to calculate each of the additions. We also asked, upon completion, to what extent they felt like they were being observed during the test (marked on a 5-point Likert scale).

Results. Based on the data (31 participants for a total of 633 additions), the average time to dismiss the debug dialogue was 1185ms and the average time to provide an answer was 9980ms. Based on these values, we conservatively consider cheating as taking more than 0.8 seconds to dismiss the spurious debug dialogue *and* taking less than 5 seconds to calculate the sum *and* providing a correct answer. It results in 147 cheating rounds (23.2% of all rounds).

Looking at these results per condition, we find 77 rounds involving cheating from 316 rounds in the human condition (24.4%) and 70 rounds involving cheating from 317 rounds in the alone condition (22.1%). A 2-samples test for equality of proportions reveals no significant difference; $\chi^2 = 0.463$, $p = .496$. This indicates no support for the presence of a human impacting the tendency to cheat. This result shows that participants do cheat relatively often, however the presence of a human observer does not significantly impact the cheating behaviour of the participants, providing no support for H1.

In terms of performance, participants in the human presence condition gave 28 wrong answers out of 239 rounds with no cheating (11.7% were wrong answers), while participants in the alone condition gave 25 wrong answers out of 247 (10.1%). Using a 2-samples test for equality of proportions, we obtain: $\chi^2 = 0.096$, $p =$

.757 indicating no support that the presence of a human impacts the performance in the test. Again, there is no significant performance difference between the two conditions, providing no support for H2. Therefore, neither of our hypotheses are supported. Due to the absence of any effects between the human and alone conditions, we did not pursue the study with robots.

Participants were asked how observed they felt on a in a Likert scale (1: "Not at all", 5: "Very much"). In the **alone** condition they felt more observed ($M = 2.69$, $SD = 1.35$) than the participants in the human presence condition ($M = 1.75$, $SD = 1.06$). A 2-tailed independent 2-samples test with equal variance assumption shows that there is significant difference: $t(28) = 2.179$, $p = .038$.

3 DISCUSSION

We can only speculate about which factors might explain our failure to observe any effect of social facilitation: the small effect sizes of social facilitation, the setting in which we collected our experimental data, or a bias towards publishing only positive results [31] that might mask how brittle social facilitation effects really are.

Participants also reported the they felt observed: in the alone condition this was $M = 2.69, SD = 1.31$, while in the human condition $M = 1.75, SD = 1.03$. When no experimenter was in the room, they felt *more* observed then when there was an experimenter in the room. This is a very notable result warranting further exploration.

The challenges of observing social interaction. What does this failed attempt at reproducing a "classic" result of social psychology tell us? Beyond possible experimental confounds, our failure at reproducing these results is likely due to the small effect size of social facilitation. In their meta-analysis of studies on social facilitation, Bond and Titus [8] showed that the overall mean effect sizes are low, ranging from 0.03 to 0.36. Uziel [38] reports weighted average effect sizes of less than 0.2. According to Cohen [9], an effect size of 0.2 should be regarded as small, an effect size of 0.5 as medium, and 0.8 as large.

Social facilitation or inhibition, like many other psychological effects, may be affected by a combination of several other factors: the observer effect (also known as the Hawthorne effect [28]), demand characteristics, cultural differences and personality. These effects are potential confounds, and adequately accounting for each of these in the experimental design is problematic.

One likely explanation is that subjects felt observed in both conditions, irrespective of a human observer sitting with them in the room. Just the process of taking part in a study might already exert a large degree of social facilitation, which is not measurably weakened or strengthened by the absence or presence of an observer in the experiment room.

The study of Guerin [18] is relevant in this context: it tried to separate the effect of observer presence from evaluation apprehension. For this a letter copying task was used in four conditions: alone; with a confederate sitting in front of the subject, but facing away; with a confederate at a desk that is behind the subject; and with a confederate sitting behind the subject with no desk in between. Guerin's results showed that there were no significant differences of errors in copying (*quality*) in any conditions, however, alone and front conditions combined were significantly different from the behind and behind-desk conditions combined in terms of task performance (*quantity*).

Furthermore, he used self-reports for determining the level of pressure the subjects felt. Subjects in the alone condition were asked to imagine how they would feel if there was a person in the room. The results showed that the subjects in the alone condition felt *more* disturbed and evaluated than those of the other three conditions, which concurs with the results we found. However, he noted that self-reports in social facilitation research may be affected by demand characteristics and self-presentation. As a result of the study, he was unable to separate evaluation apprehension from the mere presence effect on task performance.

It is likely that the subjects in our study felt observed by taking part in a study. Even though the true intent of the study was not revealed until the debriefing, subjects felt observed whatever the condition and this might have impacted their behaviour. This is know as the Hawthorne effect. However, the Hawthorne effect itself is a subject of discussion as there are studies that challenge its existence. Jones [21] studied the original experiment data [28], and found that there is slight or no evidence of a Hawthorne effect. McCambridge et al. [24] reviewed over 19 studies that investigate the Hawthorne effect, and argued that the term is used to describe a broad range of effects in the literature rather than the core definition which refers to the change in subjects' behaviour due to conformity to perceived norms or researcher expectations. Hence, they could not confirm whether the effect exists.

Weak methods in older psychology literature. Beyond the caution that must be observed when studying one specific psychological effect, a broader range of methodological issues with older research in psychology might explain why some results in psychology are incorrectly believed to be reliable.

For instance, the Bond and Titus [8] meta-analysis of research on social facilitation claims to have exhaustively examined every publication prior to the publication of the meta-analysis itself (in 1983). As a matter of fact, the oldest study that they reference dates from 1898, and 35 out of the 241 were published prior to 1965. As such, social facilitation is a good example of an old, classical psychological effect. It however also hints at the fact that its characterisation might have relied on weak research methodologies by today's standards. In that regard, Bond and Titus raise interesting points: only 100 out of the 241 studies state that the experimenter was in a different room in the alone condition (and in 96 studies, we know the experimenter was in the room). This would be seen today as a serious confound. Similarly, Bond and Titus report that 72.3% of the total participants were undergrad students, pointing to a possible demographic bias.

Biases in scientific publishing: the 'file drawer' problem. Coined in 1979 by Rosenthal [30], the *file drawer problem* refers to the bias introduced into the scientific literature by mainly publishing positive results, and rarely negative or non-confirmatory results. As a consequence, an effect could be reported and believed reliable, simply for the lack of literature showing the contrary. Rosenthal proposes to account for this problem by reporting in meta-analysis the 'fail-safe N' measure: N is the number of null effects that would be required to make the original result non-significant. Rosenthal

considers an effect resistant to the 'file drawer problem' of unreported null effects if the fail-safe N is above $5k + 10$, with k the number of reported effects.

Bond and Titus [8] report the fail-safe N for some of the effects of social facilitation. For instance, their meta-analysis show that the performance quantity of participants for complex tasks reliably decreases in presence of an observer (even thought the effect size is small). 54 effects are reported, and they note that the fail-safe N value is 160: 160 is clearly smaller than $5 \times 54 + 10 = 280$ and as such, this result could well be subject to the problem of unreported null effect. The fact that social presence inhibits the performance in complex tasks is not a robust result in the face of the bias towards publishing only positive results.

A weighted calculation of the fail-safe number has been proposed [29] that addresses some of the concerns with Rosenthal's proposal, and while not systematically reported in the literature, this metric is a valuable tool for HRI researchers when assessing how robust a result in psychology is.

4 CONCLUSION

While we have built this paper around social facilitation and our failed attempt at replicating this well-established effect, the observations we make above are broadly applicable to Human-Robot Interaction. Our failure to replicate a result from social psychology which has stood for 120 years [37] should form a cautionary tale. The limited reproducibility of results in psychology seems to be endemic [1] and while the reasons for the lack of reproducibility are many and diverse, there is a genuine concern that the field of HRI is also affected. We are, however, not suggesting that HRI should not build upon psychology anymore. Quite the contrary. Our field has strong ties with psychology, and our work is grounded in various theoretical and methodological frameworks. If anything, we encourage the community to keep on building new links with neighbouring academic fields, and social psychology should be a preferred partner in this effort.

However, we need to be frank: results from social psychology, experimental methodologies and reporting methods which were considered as commonly accepted or even gold standards until recently, are losing their special status. Instead we would like to offer the following suggestions to the HRI field:

Replicate and reproduce. When replicating a social psychology effect with robots, it is necessary to first reproduce the effect with people. Methods change, times and mores change, and negative results often go unreported. A social psychology effect which is touted in textbooks might not be that easy to replicate. With psychology at the centre of the recent replication controversy, many results which seem established should be approached with the necessary skepticism.

Null-results are interesting. The field of HRI most likely also suffers from publication bias and the file drawer effect: many studies go unreported because the results are inconclusive, negative or because they do not support an agenda. If results are negative or insignificant, the field needs to know. This helps us focus our resources better: if an experiment returned negative results and we know about it, then it can help us avoid setting up a similar experiment. It also helps us with quantifying bold claims. As results come in that are inconclusive or unsupportive of those claims, they tend to go unpublished or do not get the same amount of airtime and attention as confirmatory results. This culture should change.

Avoid questionable research practices. A number of questionable research practices (QRPs) have been identified in social psychology [20, 33]. While we have not collected data on the presence of QRPs in HRI, we need to be aware of the QRPs identified in psychology. Examples include (from [20]) selective reporting of data, or only reporting data which support a particular story; collecting data until the results are significant; p-value rounding, i.e., rounding p-values down to .05 to suggest statistical significance (a particular problem of null-hypothesis testing); failing to report all conditions; or selectively reporting studies that "worked".

Register your study. In clinical studies, it is customary to register the study protocol before beginning data collection (see for example clinicaltrials.gov). Perhaps a similar practice should be established for HRI. Among the many benefits, the registration of trials before running include the reduction of publication bias, the efficient allocation of research resources, and full engagement with ethical obligations of the research community.

Avoid the Hawthorne effect. The set-up of most HRI studies often reveals to the subjects that they are being observed: lab-based studies always implicitly signal to subjects that their behaviour will be monitored. Even moving into a naturalistic environment might not alleviate this problem, as ethics procedures insist that subjects are briefed before a study and that their explicit consent is sought before they can engage in the experiment. As such, subjects in HRI experiments might always experience the Hawthorne effect: their behaviour changes because they are aware of being observed. The only way forward here is to either not inform subjects prior to the study (which is unethical) or work with a distractor task. However, the latter is particularly difficult to implement in HRI.

Come up with HRI reference tasks. While there is merit in attempting to reproduce effects from social psychology with robots instead of people, it might be worth identifying new effects and tasks relevant to Human-Robot Interaction and its applications. Times change and as robots become more ubiquitous, our response to robots is likely to evolve rapidly. We need to look at the relation and interaction between people and robots through new lenses, and the old (often very old) views from social psychology are perhaps no longer applicable or appropriate. It may be noted that our implementation of the methodology did not perfectly match one from psychology. The task used in the second attempt was the same as one from psychology, however, we did not deploy the essay writing portion of the original [40] so as not to introduce a confound. Finding an appropriate methodology to replicate in the context of HRI was a challenge in itself, further reinforcing the need for our own reference tasks.

As a community, HRI should learn from its own mistakes (see Baxter et al. [6] for good advice) and from the mistakes of others. We are a young community, with a steady influx of young talent, and we often look towards established fields for guidance. But when exactly these established fields start to question their own practices

and results, we should too. The conclusion of the Science study on reproducibility in psychology [1] offers the following message:

> Following this intensive effort to reproduce a sample of published psychological findings, how many of the effects can we confirm are true? Zero. And, how many of the effects can we confirm are false? Zero. Is this a limitation of the project design? No. It is the reality of doing science, even if it is not appreciated in daily practice.

Importantly, this is the reality of doing science *in general*, not only *social* science. We must not blind ourselves: our methods and protocols in HRI do not shelter us from the exact same problems experienced in other fields. Future researchers may well write the same kind of article about our field when they revisit today's literature on Human-Robot Interaction.

ACKNOWLEDGMENTS

This work has been partially supported by the EU H2020 L2TOR project (grant 688014), the EU FP7 DREAM project (grant 611391), the EU H2020 Marie Sklodowska-Curie Actions Innovative Training Networks project APRIL (grant 674868), and the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227). All authors have contributed equally to the experimental design, execution, data analyses and writing.

REFERENCES

[1] A.A. Aarts, C.J. Anderson, J. Anderson, M.A.L.M. van Assen, P.R. Attridge, et al. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015). https://doi.org/10.1126/science.aac4716

[2] F.H. Allport. 1924. *Social Psychology.* Houghton Mifflin Company, Chapter Response to social stimulation in the group, 260–291.

[3] M. Baker. 2016. 1500 scientists lift the lid on reproducibility. *Nature News* 533, 7604 (2016), 452. https://doi.org/10.1038/533452a

[4] C. Bartneck. 2011. The End of the Beginning: A Reflection on the First Five Years of the HRI Conference. *Scientometrics* 86, 2 (2011), 487–504.

[5] C. Bartneck. 2017. Reviewers' scores do not predict impact - Bibliometric Analysis of the Proceedings of the Human-Robot Interaction Conference. *Scientometrics* 110, 1 (2017), 179–194. https://doi.org/10.1007/s11192-016-2176-y

[6] P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme. 2016. From characterising three years of HRI to methodology and reporting recommendations. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 391–398.

[7] C.F. Bond. 1982. Social Facilitation: A self-presentational view. *Journal of Personality and Social Psychology* 42, 6 (1982), 1042–1050.

[8] C.F. Bond and L.J. Titus. 1983. Social facilitation: a meta-analysis of 241 studies. *Psychological bulletin* 94, 2 (1983), 265–292. https://doi.org/10.1037/0033-2909.94.2.265

[9] J. Cohen. 1977. *Statistical power analysis for the behavioral sciences*. Academic Press.

[10] N.B. Cottrell, R.H. Rittle, and D.L. Wack. 1967. The presence of an audience and list type (competitional or noncompetitional) as joint determinants of performance in paired associates learning. *Journal of Personality* 35 (1967), 425–434.

[11] W.D. Criddle. [n. d.]. The physical presence of other individuals as factor in social facilitation. *Psychonomic Science* 22, 4 ([n. d.]), 229–230.

[12] J.M. Feinberg and J.R. Aiello. 2010. The Effect of Challenge and Threat Appraisals Under Evaluative Presence. *Journal of Applied Social Psychology* 40, 8 (2010), 2071–2104.

[13] T.R. Fosgaard, L.G. Hansen, and M. Piovesan. 2013. Separating Will from Grace: An experiment on conformity and awareness in cheating. *Journal of Economic Behavior and Organization* 93 (2013), 279–284. https://doi.org/10.1016/j.jebo.2013.03.027

[14] V.J. Ganzer. 1968. Effects of audience presence and test anxiety on learning and retention in a serial learning situation. *Journal of Personality and Social Psychology* 8, 2 (Pt. 1) (1968), 194–199.

[15] W.L. Gardner and M.L. Knowles. 2008. Love Makes You Real: Favorite Television Characters Are Perceived as "Real" in a Social Facilitation Paradigm. *Social Cognition* 26, 2 (2008), 156–168. https://doi.org/10.1521/soco.2008.26.2.156

[16] R.G. Geen. 1973. Effects of being observed on short- and long-term recall. *Journal of Experimental Psychology* 100 (1973), 395–398.

[17] B. Guerin. 1983. Social Facilitation and social monitoring: a test of three models. *British Journal of Social Psychology* 22 (1983), 203–214. https://doi.org/10.1111/j.2044-8309.1983.tb00585.x

[18] B. Guerin. 1989. Reducing evaluation effects in mere presence. *The Journal of Social Psychology* 129, 2 (1989), 183–190.

[19] J.P. Hill and R.A. Kochendorfer. 1969. Knowledge of peer success and risk of detection as determinants of cheating. *Developmental Psychology* 1, 3 (1969), 231–238. https://doi.org/10.1037/h0027330

[20] L.K. John, G. Loewenstein, and D. Prelec. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* 23, 5 (2012), 524–532.

[21] S.R.G. Jones. 1992. Was There a Hawthorne Effect? *Amer. J. Sociology* 98, 3 (1992), 451–468.

[22] J.P. Lombardo and J.F. Catalano. 1975. The effect of failure and the nature of the audience on performance of a complex motor task. *Journal of Motor Behavior* 7 (1975), 29–35.

[23] R.J. McCaffrey, J.M. Fisher, B.A. Gold, and J.K. Lynch. 1996. Presence of third parties during neuropsychological evaluations: Who is evaluating whom? *The Clinical Neuropsychologist* 10, 4 (1996), 435–449. https://doi.org/10.1080/13854049608406704

[24] J. McCambridge, J. Witton, and D.R. Elbourne. 2014. Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology* 67, 3 (2014), 267–277.

[25] F.G. Miller, M.E Hurkman, J.B. Robinson, and R.A. Feinberg. 1979. Status and evaluation potential in the social facilitation and impairment of task performance. *Personality and Social Psychology Bulletin* 5 (1979), 381–385.

[26] D.S. Nagin and G. Pogarsky. 2003. An Experimental Investigation of Deterrence: Cheating, Self-Serving Bias, and Impulsivity. *Criminology* 41, 1 (2003), 167–194. https://doi.org/10.1111/j.1745-9125.2003.tb00985.x

[27] N. Riether, F. Hegel, B. Wrede, and G. Horstmann. 2012. Social facilitation with social robots?. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12*. 41–47. https://doi.org/10.1145/2157689.2157697

[28] F.J. Roethlisberger, W.J. Dickson, H.A. Wright, and Western Electric Company. 1939. *Management and the Worker: An Account of a Research Program Conducted by the Western Electric Company, Hawthorne Works, Chicago.* Harvard University Press.

[29] M.S. Rosenberg. 2005. The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution* 59, 2 (2005), 464–468.

[30] R. Rosenthal. 1979. The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86, 3 (1979), 638–641.

[31] H.R. Rothstein, A.J. Sutton, and M. Borenstein. 2006. *Publication bias in meta-analysis: Prevention, assessment and adjustments.* John Wiley & Sons.

[32] P. Schermerhorn, M. Scheutz, and C.R. Crowell. 2008. Robot Social Presence and Gender: Do Females View Robots Differently than Males? *ACM/IEEE International Conference on Human-Robot Interaction* (2008), 263–270. https://doi.org/10.1145/1349822.1349857

[33] J.P. Simmons, L.D. Nelson, and U. Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* 22, 11 (2011), 1359–1366.

[34] W. Stroebe. 2012. The truth about Triplett (1898), but nobody seems to care. *Perspectives on Psychological Science* 7, 1 (2012), 54–57.

[35] M.J. Strube, M.E. Miles, and W.H. Finch. 1981. The social facilitation of a simple task: Field tests of alternative explanations. *Personality and Social Psychology Bulletin* 7 (1981), 701–707.

[36] D.J. Terry and M. Kearnes. 1993. Effects of an audience on the task performance of subjects with high and low self-esteem. *Personality and Individual Differences* 15, 2 (1993), 137–145.

[37] N. Triplett. 1898. The dynamogenic factors in pacemaking and competition. *American Journal of Psychology* 9, 4 (1898), 507–533.

[38] L. Uziel. 2007. Individual differences in the social facilitation effect: A review and meta-analysis. *Journal of Research in Personality* 41, 3 (2007), 579–601. https://doi.org/10.1016/j.jrp.2006.06.008

[39] F.T. Vitro and L.A. Schoer. 1972. The effects of probability of test success, test importance, and risk of detection on the incidence of cheating. *Journal of School Psychology* 10, 3 (1972), 269–277. https://doi.org/10.1016/0022-4405(72)90062-3

[40] K.D. Vohs and W. Schooler. 2008. The Value of Believing in Free Will: Encouraging a Belief in Determinism Increases Cheating. *Psychological Science* 19, 1 (2008), 49–54. https://doi.org/10.1111/j.1467-9280.2008.02045.x

[41] I. Wechsung, P. Ehrenbrink, R. Schleicher, and S. Möller. 2014. Investigating the Social Facilitation Effect in Human-Robot Interaction. In *Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialog Systems into Practice*, J. Mariani, S. Rosset, M. Garnier-Rizet, and L. Devillers (Eds.). Springer, New York, 167–177.

[42] R.B. Zajonc. 1965. Social facilitation. *Science* 149, 3681 (1965), 269–274.

Multi-modal Open-Set Person Identification in HRI

Bahar Irfan
CRNS, Plymouth University, UK
bahar.irfan@plymouth.ac.uk

Natalia Lyubova
SoftBank Robotics Europe, France
nlyubova@softbankrobotics.com

Michael Garcia Ortiz
SoftBank Robotics Europe, France
mgarciaortiz@softbankrobotics.com

Tony Belpaeme
IDLab - imec, Ghent University, Belgium
CRNS, Plymouth University, UK
tony.belpaeme@plymouth.ac.uk

ABSTRACT

In this paper, we describe a multi-modal Bayesian network for person recognition in a HRI context, combining information about a person's face, gender, age, and height estimates, with the time of interaction. We conduct an initial study with 14 participants over a four-week period to validate the system and learn the optimal weights for each of the metrics. Several normalisation methods are compared for different settings, such as learning from data, face recognition threshold and quality of the estimation. The results show that the proposed network improves the overall recognition rate by at least 1.4% comparing to person recognition based on face only in an open-set identification problem, and at least 4.4% in a closed-set.

KEYWORDS

Person recognition; Bayesian network; multi-modal data fusion; soft biometrics; personalisation

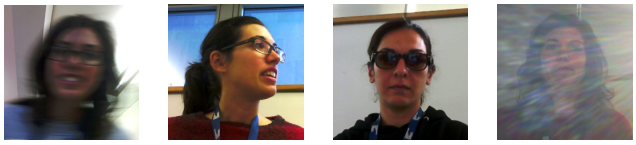
1 INTRODUCTION

Recognising a person is an essential step in establishing a personalised long-term human-robot interaction (HRI). In contrast to verification problems, where a user would state her identity and the system confirms or rejects it, in an HRI scenario, automatic recognition is desired for a natural interaction. In addition, the user might not be encountered before, in which case, the robot is expected to "meet" the user, i.e. enroll the user into the system. This problem is classified as an *open-set identification problem*, which is more difficult than closed-set identification or verification problems [5].

Biometric systems generally perform user recognition based on face recognition (FR). However, most FR challenges such as Face Recognition Vendor Tests[1] evaluate algorithms that perform verification. To this date, the only available open-set identification challenge is the Unconstrained Face Detection and Open Set Recognition Challenge[2], which shows that the algorithms achieve good identification accuracies at high false identification rates [6].

[1]https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt
[2]http://vast.uccs.edu/Opensetface

HRI '18 Workshop, March 5-8, 2018, Chicago, IL, USA



(a) (b) (c) (d)

Figure 1: Examples of unreliable face recognition from our study: (a) a blurry image; (b) an oblique viewing angle; (c) occlusions, e.g. glasses; (d) lighting condition, e.g. direct light.

Similarly, during a real-time interaction, FR could be unreliable due to a number of reasons including changing facial features, expressions, and lighting conditions [16] (see Fig. 1). Another example is the recent release of a smart phone with the built-in FR system for unlocking the phone that struggle to distinguish family members due to similarity of their facial features [4]. This issue increased awareness of the security and privacy problems that might arise from using a uni-modal biometric system that might not be as reliable as using a pass code.

Moreover, a biometric system may not be able to obtain meaningful data in some cases, resulting in a *failure-to-enroll* (FTE) error [13]. For instance, a face may not be detected in a blurry image where a person is moving. In addition, the upper bound on identification accuracy would limit the matching performance of a uni-modal biometric system. However, multi-modal biometric systems can improve the matching accuracy of a recognition by fusing information from multiple sources that could reduce the effects of noisy data, decrease FTE error, and eliminate the upper bound issue for a better determination of the identity. Robots, due to the rich sensor suite they carry, lend themselves well to multi-modal recognition.

In this paper, we explore a multi-modal Bayesian network (BN) for integrating soft biometric information, such as a user's gender, age, height, and time of interaction, together with the primary biometric information provided by face recognition. These biometric modalities are non-intrusive and can be obtained using the camera embedded on the robot. We designed a pilot study to validate our system in a real-time HRI scenario. We compare performances of several normalisation methods using optimised weights for each comparison. The proposed recognition system is intended to be used in a real-world application in Cardiac Rehabilitation therapy with a personalised robot [10].

2 RELATED WORK

Several post-classification fusion methods have been proposed for integrating multi-modal information in biometrics. They can be classified into three main categories: decision, rank, and confidence level fusion [8]. Decision level fusion (e.g. majority voting, AND/OR rule) is based on combining individual best matches from each biometric matcher. Rank level methods (e.g. highest rank, logistic regression) are used when the output of each biometric matcher consists of ranked matches. Confidence level fusion is the most common approach, as it allows a weighted decision from multiple biometric classifiers. There are two main approaches for combining the scores for confidence level fusion: classification methods and combination approaches. Classification methods treat the classifications of individual classifiers as input for a new classifier, such as ANNs and Support Vector Machines, which allows combining non-homogeneous data without preprocessing. The combination approach on the other hand consist of three steps: (1) normalisation of scores from different modalities into a common domain, (2) combination of scores through a method such as *sum* or *product rule*, and (3) thresholding to obtain the identification results. Performances of the combination approaches depend on the method and threshold chosen at each of the steps.

Although most biometric systems utilise primary biometrics for person recognition, such as fingerprint or face, other attributes of an individual such as age, gender, and clothing –referred to as soft biometrics– can provide additional information to improve the recognition performance [2]. In [9], the authors proposed combining a primary biometric trait (fingerprint) with soft biometric traits, such as a person's gender, ethnicity, and height, using a BN. In the weighting scheme, the traits with smaller variability and larger distinguishing capability were given more weight in the computation of the final matching probabilities. Furthermore, smaller weights were assigned to the soft biometric traits, so that if a soft biometric trait is measured incorrectly (e.g. a male user being identified as a female) the rejection probability is decreased. They achieved a 4% improvement in the genuine acceptance rate, however, the fusion weights were not optimised.

To the best of our knowledge, our approach is the first in combining soft biometrics with a primary biometric to identify a user in real-time, in the field of HRI. Moreover, it is the first time that the presented modalities (face, gender, age, height and time of interaction) are fused together, although they have shown improvement when fused separately or with other biometrics [2, 11, 15].

3 METHODOLOGY

We developed a BN based on [9] integrating multi-modal biometrics for reliable recognition in real-time human-robot interaction. We fuse face recognition (F) information (primary biometric) with gender (G), age (A), and height (H) estimations and the time of the interaction (T) (soft biometrics). Conditional independence is assumed between nodes, given the identity (I). The pyAgrum [3] library is used for implementing the structure.

3.1 Structure

The states of each node are determined by: the number of known users (for F and I), the available range of the modality (for A and H), and the pre-defined values (for G, "female" and "male", and for T, the day of the week combined with the time of the interaction). The data available about each user are converted to probabilities for each state within a node. These values are used as evidence in the network, and the maximum posterior for the I node determines the estimated identity of the user.

FR values are assumed to be similarity scores, such that, each score gives the percentage of similarity of the current user to the faces in the database. These scores are normalised to find the probabilities of the states. Age, height, and time are considered as discrete random variables (e.g. age is taken as 26, between 26 and 27). We estimate the probabilities of the remaining states by assuming a discretised and normalised normal distribution, $N(\mu, \sigma^2)$, defined by Eq. 1, where X is the estimated value, Z is the z-score, and C is the confidence of the biometric indicator for the estimated value.

$$\mu = X, \quad P(\frac{-0.5}{\sigma} < Z < \frac{0.5}{\sigma}) = C \tag{1}$$

Generally, in FR systems, if the highest similarity score or probability is below a given threshold, the user is declared as "unknown". However, in a BN, the posterior probabilities can be quite low due to the multiplication of probabilities during inference and the value can decrease with increasing number of states in a node. Thus, instead of using a fixed threshold in our system, we use the quality of the estimation (Q), in which we compare the highest probability (P_w) to the second highest probability (P_s), as shown in Eq. 2. The difference is multiplied by the number of people in the database (n_p), because the difference between the probabilities decreases as the number of people increases (the sum of probabilities is 1.0). Initially $Q = 0$, which eliminates the cases where the first and the second highest probabilities are the same.

$$Q = [P_w(I|F, G, A, H, T) - P_s(I|F, G, A, H, T)] * n_p \tag{2}$$

The FR threshold (θ_{FR}) is maintained in the system through the introduction of the "unknown" (U) state in face and identity nodes. The similarity score of U for FR is set to the θ_{FR}, hence, when normalised, the similarity scores below the threshold have lower probabilities than U. Similarly, those that have higher similarity scores than the threshold will have higher probabilities than U.

3.2 Learning

Our hypothesis is that the recognition could be improved by learning the likelihoods of the system through evidence. Hence, as our contribution, we propose a BN where the likelihoods of the system are learned from data.

A possible solution could be to create a model that depends on time-series data, like a dynamic BN. However, in a dynamic BN, only the immediate prior value at the previous time step is used, which differs from an open-set identification problem, where the previous state can contain values that applied to another user. For example, user "1" might be encountered right before user "2"; in which case, the evidence for user "1" might not have an effect on the identity estimation for user "2".

Therefore, we designed our own approach in learning from data. We initially use the prior knowledge in setting the likelihoods for each variable (e.g. $P(F|I)$, $P(G|I)$). When the robot identifies a user, and the user confirms her identity, the recognition information and the identity of the user are fed as evidence to the network, and the

current posteriors are summed and normalised with the previous posteriors, to update the posteriors for this user. In our example (see Fig. 2), initially, $P(F =\text{"1"}|I =\text{"1"})$ is set to be much greater than the rest of the likelihoods. However, the FR evidence gives a higher probability score for "3" than "1", which might be due to the similarity in their appearance. After the identity confirmation of the user, using the face evidence, and the evidence for the other modalities, the likelihood is updated by summing with the previous posterior and then normalising it. Updating the posteriors would allow the network to learn their similarity, hence, at the next encounter, the probability for mistaking "1" with "3" would be decreased.

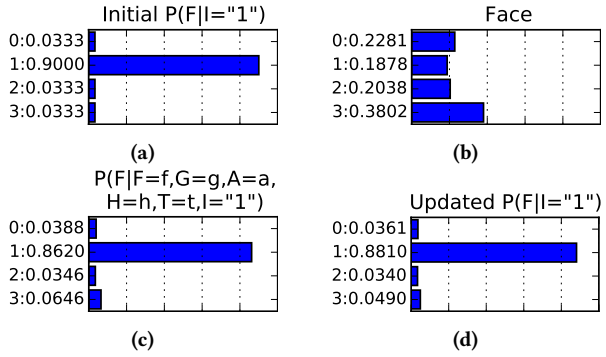


Figure 2: Learning: (a) Initial likelihood of F given I = "1", (b) F evidence, (c) posterior using the evidence, (d) updated posterior.

Likewise, if the recognised user was not previously enrolled into the system, then posterior of $P(F|I =\text{"0"})$ is updated. However, gender, age, height, and time posteriors for U are not changed, as they should be uniformly distributed. In order to allow the network to learn from enough data to make meaningful estimations, the output of the system is returned as U if the number of recognitions is less than a predetermined threshold (here, we chose 5).

3.3 Weights

We smooth the recognition results of each modality by using the weights as an exponent to the results for the evidence, due to using product rule as the combination method, as opposed to the sum of logarithms method in [9]. Also, we do not restrict the sum of weights to 1.0, as this could deteriorate results of the primary biometric trait (face), and instead set the weights to the range from 0.0 to 1.0.

We designed a pilot experiment, described in Section 4, for collecting data to optimise the weights, that would minimise the overall recognition error. The weights were optimised for each parameter separately, except for the weight of F, which is always 1.0. The weights that corresponded to the minimum number of incorrect recognitions were combined to get the optimum weights, based on the assumption that each node is conditionally independent.

3.4 Normalisation

A good normalisation method should be insensitive to the outliers and provide a good estimate of the real distribution [8]. For analysing the effects on the performance, we compared such normalisation methods that scale the values to $[0, 1]$ range to be used as probabilities within the BN: min-max, tanh [7], softmax [1], and norm-sum (dividing by the sum of values).

BNs use the *product rule* for combining the results of each node, hence, if a probability of a classifier is zero, it results in an overall zero probability for a class irrespective of the results from other classifiers. In order to overcome this problem we used a small cut-off probability threshold as $p_t = 10^{-6}$.

3.5 Extendability

Our approach relies on FR primarily, but the described system can be extended with other primary biometric traits such as voice and fingerprint, and soft biometric traits, such as the location of the interaction, and ethnicity. It is intended to increase the recognition rate from a single image, and tracking is not applied between images. In order to increase the reliability of the system, multiple images (3 images here) are taken in succession during the pilot study and the results are normalised to estimate the identity of the user, which allows discarding the images without a face detected.

The system does not require heavy-computing, hence, it is suitable for use on commercially available robots. We use a Pepper robot[3] in our study with Naoqi[4] software modules (providing a user's face ID, gender, age, and height that we used as input modalities), however, the network is applicable to any recognition software.

4 STUDY IN USER IDENTIFICATION

The objective of this study is to gather data for finding the optimal weights for the proposed network for user identification.

4.1 Protocol

The user initially enrolls to the system by entering his/her name, gender, age, and height, and then the robot takes photos of the user. During next encounters, the robot predicts the identity of the user and asks for confirmation.

We ran the study with 14 participants (4 female, 10 male, of age range 24-40) and collected a total of 66 images per user over the four weeks period. The recognition process took approximately 5 seconds: ~2-3s for user detection, ~1.5s (0.5s each) for image capture, ~1s to load the network parameters, ~0.6s (0.2s each) for recognition from modalities, ~0.9s (0.3s each) for estimation of the identity using the network. The robot stayed in a fixed position before the interaction, and only when a user was identified, it would become animate to ensure a natural interaction. We aimed to achieve better quality of images by keeping the robot fixed, however, since the robot did not notify participants when taking images, some of the captured images include people looking sideways, smiling, partially covering their faces or moving (see Fig. 1).

We use single-user recognition within the images, that is, only one user is assumed to be present in front of the camera. Hence, the image database was cleaned of images with multiple people or any other user rather than the claimed identity for cross-validation. However, in the future, the position of each user can be considered for multiple people recognition and interaction.

[3] https://www.ald.softbankrobotics.com/en/robots/pepper
[4] http://doc.aldebaran.com/2-5

4.2 Results

In order to validate our system, 5-fold cross-validation is applied with 13 images per user in each bin with a different randomised initial ordering of the users, and the results are averaged. Detection and identification rates (DIR) and false alarm rates (FAR) are reported for the pilot study along with receiver operating characteristics (ROC) curves (see Appendix and Fig. 3), which are the performance measures for the open-set identification problem [12].

The average failure to enroll error (FTE) is 0.214 (0.008), which corresponds to the fraction of images where a face cannot be detected. The identity was not estimated by the network in those cases because the only primary biometric in our system is FR and soft biometrics do not have the deterministic characteristic to estimate the identity on their own.

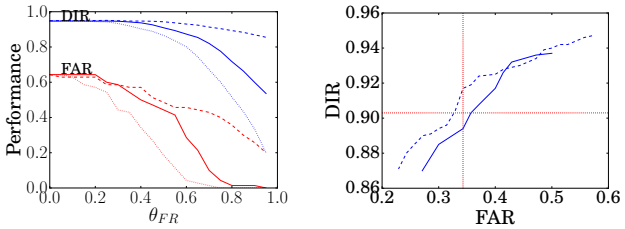
The optimised weights (see Appendix) show that in our study the age is the least effective soft biometric in determining the identity, whereas height is the most effective one. However, this might be due to the characteristics of the population in our pilot study, as the participants' ages are close to each other. Another important factor is the reliability of the age recognition software. The standard deviation of the estimated age of a user on average was 9.3. Hence, we cannot conclude that age should not be used to supplement the FR in general, but if used, the accuracy of the software used should be high, especially in a population with a narrow age range. On the other hand, the effectiveness of the height can also be explained by the nature of the population (3 relatively tall (> 180 cm) and 2 relatively short (< 160 cm) users), even though the average standard deviation of estimated height was 6.3 cm. A more balanced dataset would allow observing the true effects of these parameters.

The cross-validation results for the optimised weights (see Appendix) show that combining soft biometrics using our proposed BN can increase the DIR, depending on the inner settings. It can be observed that although norm-sum and min-max methods provide good results without learning, the recognition rate drops below FR with learning, whereas softmax and tanh methods are not affected.

However, the FAR of the network for any normalisation method is greater than the FAR of FR. This is caused by the combination of multimodal data. For example, if the highest face similarity score is below the threshold, the FR reports the user as "unknown". The network, on the other hand, will still try to identify the user based on other sensor input, where errors might increase FAR.

In order to compare the effects of learning, we chose the min-max method without learning and with a cut-off threshold (N_{minmax}) and the softmax method with learning and no cut-off threshold ($NL_{softmax}$), because the former provides the second highest DIR but with lower FAR than that of the best methods (highlighted in blue), and the latter provides the best DIR in learning in both training and test sets. The results are presented in Fig. 3.

The trade-off between DIR and FAR can be observed in Fig. 3a. The ideal FR threshold (θ_{FR}) should maintain a low FAR with a good DIR. For example, at $\theta_{FR} = 0.7$, the FAR is very low for both FR and $NL_{softmax}$, however, the DIR has also decreased substantially. If we compare the results in the range where $FAR_{FR} \leq 0.5$ ($\theta_{FR} = 0.3$) and $DIR_{FR} \geq 0.8$ ($\theta_{FR} = 0.6$): N_{minmax} is better in identification ($0.93 \leq DIR_{minmax} \leq 0.949$) than $NL_{softmax}$ ($0.873 \leq DIR_{softmax} \leq 0.946$) and FR ($0.801 \leq DIR_{FR} \leq 0.933$). However, $NL_{softmax}$ misidentifies



(a) (b)

Figure 3: ROC curves (Dotted lines represent FR results, dashed line is N_{minmax}, solid line is $NL_{softmax}$): (a) Performance measures, DIR (in blue) and FAR (in red), for varying θ_{FR}; (b) ROC curve for varying Q values for $\theta_{FR} = 0.4$.

the "unknown" users much less ($0.286 \leq FAR_{softmax} \leq 0.543$) than N_{minmax} ($0.457 \leq FAR_{minmax} \leq 0.571$).

$\theta_{FR} = 0.4$ gives the highest detection rate for both N_{minmax} and $NL_{softmax}$ with lower FAR, hence, we compared the effects of quality of estimation (Q) at this rate (see Fig. 3b). The area of improvement for the open-set identification problem is where $FAR \leq FAR_{FR}$ and $DIR \geq DIR_{FR}$. $NL_{softmax}$ does not provide a value in this range, hence, we can conclude that the proposed learning method performs worse than the method without learning. DIR_{minmax} is 1.4% higher than DIR_{FR} where the FAR is equal ($Q = 0.31$), and FAR_{minmax} is 1.4% lower than FAR_{FR} where DIR is equal ($Q = 0.41$). On the other hand, if the problem was treated as a closed-set problem (where all the users are enrolled into the system), $Q = 0$ would be sufficient and the increase in DIR would be 4.4%.

5 CONCLUSION

Our results suggest that the use of soft biometrics increases the recognition rate, however, it can also increase the misidentification rate of unknown users. Increasing θ_{FR} and Q can indeed decrease the FAR, but it can decrease the DIR as well. On the other hand, our proposed learning method mostly performs worse than the traditional BN on this dataset.

Furthermore, the results indicated that our dataset might be biased due to the small population size and the characteristics of the population. To our knowledge, the only publicly available database that contains the soft biometrics used in our system (except the time of interaction) with a face database is the recently released BioSoft [14]. However, the number of subjects is limited to 75, and the height is defined in labels instead of numeric values. Therefore, as a future extension, we will generate an artificial database with a higher amount of subjects with differing soft biometrics, which would also allow setting the noise level in the modalities. We aim to compare the performance of our system with other classification methods such as Support Vector Machines on the artificial dataset.

In the near future, we plan to use the proposed user recognition system in Cardiac Rehabilitation (CR) therapy, during which the robot will recognise the patients and personalise the interaction based on the information about the patients' previous sessions and their progress during the therapy [10]. The study will allow us to evaluate our system in a real-world application, and to observe the effects of personalisation in a long-term HRI.

Appendix: 5-fold cross validation mean (with standard deviation) of false alarm rates (FAR) on the training set, detection and identification rates (DIR) for rank 1 for training and test sets, and optimised weights for each normalisation method with varying learning method and cut-off threshold (p_t) settings with $\theta_{FR} = 0.3$. Highlights in blue show the best values obtained in learning and without learning conditions (minimum FAR, maximum DIR for training and test sets). Highlights in red show the chosen methods for the comparison of learning from data.

Learning	p_t	*Normalisation*	*FAR*	DIR_1 *(Training)*	DIR_1 *(Test)*	w_G	w_A	w_H	w_T
none	none	FR	0.443 (0.078)	0.933 (0.004)	0.945 (0.015)	0	0	0	0
none	none	norm-sum	0.629 (0.032)	0.951 (0.004)	0.967 (0.013)	0	0	0.1	0
none	none	min-max	0.629 (0.032)	0.951 (0.005)	0.965 (0.015)	0.2	0	0.1	0
none	none	softmax	0.571 (0.072)	0.947 (0.004)	0.965 (0.014)	0.1	0	0.6	0
none	none	tanh	0.571 (0.051)	0.942 (0.005)	0.955 (0.012)	0	0	0.1	0
none	1e-6	norm-sum	0.529 (0.081)	0.943 (0.003)	0.956 (0.015)	0	0	0.1	0.1
none	1e-6	min-max	0.586 (0.060)	0.949 (0.005)	0.965 (0.014)	0.2	0	0.1	0
none	1e-6	softmax	0.571 (0.072)	0.946 (0.003)	0.959 (0.015)	0.1	0.1	0.1	0.1
none	1e-6	tanh	0.543 (0.039)	0.942 (0.003)	0.957 (0.013)	0	0	0.3	0.1
evidence	none	norm-sum	0.629 (0.032)	0.782 (0.063)	0.694 (0.093)	0.1	0	0.1	0
evidence	none	min-max	0.629 (0.032)	0.776 (0.064)	0.692 (0.090)	0	0	0.1	0
evidence	none	softmax	0.586 (0.060)	0.946 (0.005)	0.961 (0.017)	0.1	0	0.6	0
evidence	none	tanh	0.571 (0.051)	0.943 (0.007)	0.955 (0.012)	0	0	0.1	0
evidence	1e-6	norm-sum	0.571 (0.072)	0.75 (0.082)	0.632 (0.127)	0.1	0	0.1	0
evidence	1e-6	min-max	0.643 (0.0)	0.776 (0.061)	0.697 (0.089)	0	0	0.1	0
evidence	1e-6	softmax	0.586 (0.060)	0.946 (0.005)	0.954 (0.025)	0.1	0	0.6	0
evidence	1e-6	tanh	0.543 (0.039)	0.943 (0.006)	0.961 (0.019)	0	0	0.3	0

ACKNOWLEDGMENTS

This work has been supported by the EU H2020 Marie Sklodowska-Curie Actions ITN project APRIL (grant 674868). The authors would like to thank Valerio Biscione for his constructive suggestions in the design of the Bayesian network, and the participants of the user study for their time and efforts.

REFERENCES

[1] Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc.

[2] Antitza Dantcheva, Petros Elia, and Arun Ross. 2016. What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security* 11, 3 (2016), 441–467. https://doi.org/10.1109/TIFS.2015.2480381

[3] Christophe Gonzales, Lionel Torti, and Pierre-Henri Wuillemin. 2017. aGrUM: a Graphical Universal Model framework. In *International Conference on Industrial Engineering, Other Applications of Applied Intelligent Systems (Proceedings of the 30th International Conference on Industrial Engineering, Other Applications of Applied Intelligent Systems).* https://doi.org/10.1007/978-3-319-60045-1_20

[4] Andy Greenberg. 2017. Watch a 10-year-old's face unlock his mom's iPhone X. (November 2017). Retrieved January 5, 2018 from https://www.wired.com/story/10-year-old-face-id-unlocks-mothers-iphone-x/.

[5] Manuel Günther, Steve Cruz, Ethan M. Rudd, and Terrance E. Boult. 2017. Toward open-set face recognition. (2017). arXiv:1705.01567 Retrieved from http://arxiv.org/abs/1705.01567.

[6] M. Günther, P. Hu, C. Herrmann, C. H. Chan, M. Jiang, S. Yang, A. R. Dhamija, D. Ramanan, J. Beyerer, J. Kittler, M. A. Jazaery, M. I. Nouyed, G. Guo, C. Stankiewicz, and T. E. Boult. 2017. Unconstrained face detection and open-set face recognition Challenge. (2017). arXiv:1708.02337 Retrieved from http://arxiv.org/abs/1708.02337.

[7] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions.* Wiley.

[8] Anil Jain, Karthik Nandakumar, and Arun Ross. 2005. Score normalization in multimodal biometric systems. *Pattern Recognition* 38, 12 (2005), 2270–2285. https://doi.org/10.1016/j.patcog.2005.01.012

[9] Anil K. Jain, Sarat C. Dass, and Karthik Nandakumar. 2004. Soft biometric traits for personal recognition systems. In *International Conference on Biometric Authentication (LNCS).* 731–738. https://doi.org/10.1007/978-3-540-25948-0_99

[10] Juan S. Lara, Jonathan Casas, Andres Aguirre, Marcela Munera, Monica Rincon-Roncancio, Bahar Irfan, Emmanuel Senft, Tony Belpaeme, and Carlos A. Cifuentes. 2017. Human-robot sensor interface for cardiac rehabilitation. In *2017 International Conference on Rehabilitation Robotics (ICORR).* 1013–1018. https://doi.org/10.1109/ICORR.2017.8009382

[11] Eric Martinson, Wallace Lawson, and Greg Trafton. 2013. Identifying people with soft-biometrics at fleet week. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction (HRI '13).* IEEE Press, 49–56.

[12] P. Jonathon Phillips, Patrick Grother, and Ross Micheals. 2011. Evaluation methods in face recognition. In *Handbook of Face Recognition* (2nd ed.), Stan Z. Li and Anil K. Jain (Eds.). Springer Publishing Company, Incorporated, 553–556. https://doi.org/10.1007/978-0-85729-932-1_21

[13] Arun Ross and Anil K. Jain. 2007. Human recognition using biometrics: an overview. *Annales Des Télécommunications* 62, 1 (2007), 11–35. https://doi.org/10.1007/BF03253248

[14] D. Sadhya, P. Pahariya, R. Yadav, A. Rastogi, A. Kumar, L. Sharma, and S. K. Singh. 2017. BioSoft - a multimodal biometric database incorporating soft traits. In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA).* 1–6. https://doi.org/10.1109/ISBA.2017.7947693

[15] Walter J. Scheirer, Neeraj Kumar, Karl Ricanek, Peter N. Belhumeur, and Terrance E. Boult. 2011. Fusing with context: A Bayesian approach to combining descriptive attributes. In *2011 International Joint Conference on Biometrics (IJCB).* 1–8. https://doi.org/10.1109/IJCB.2011.6117490

[16] Waldemar Wójcik, Konrad Gromaszek, and Muhtar Junisbekov. 2016. Face recognition: Issues, methods and alternative applications. In *Face Recognition - Semisupervised Classification, Subspace Projection and Evaluation Methods*, S. Ramakrishnan (Ed.). InTech, Chapter 02. https://doi.org/10.5772/61471

A Robot Teaching Young Children a Second Language: The Effect of Multiple Interactions on Engagement and Perfor....

Conference Paper · March 2018

DOI: 10.1145/3173386.3177059

CITATIONS
0

READS
5

5 authors, including:

Rianne van den Berghe
Utrecht University
5 PUBLICATIONS **1** CITATION

SEE PROFILE

A Robot Teaching Young Children a Second Language: The Effect of Multiple Interactions on Engagement and Performance

Emmy Rintjema
Tilburg University
P.O. Box 90153, 5000 LE Tilburg
The Netherlands
emmyrintjema@gmail.com

Rianne van den Berghe
Utrecht University
P.O. Box 80125, 3508 TC Utrecht
The Netherlands
m.a.j.vandenberghe@uu.nl

Anne Kessels
Tilburg University
P.O. Box 90153, 5000 LE Tilburg
The Netherlands
kesselsanne@gmail.com

Jan de Wit
Tilburg University
P.O. Box 90153, 5000 LE Tilburg
The Netherlands
j.m.s.dewit@uvt.nl

Paul Vogt
Tilburg University
P.O. Box 90153, 5000 LE Tilburg
The Netherlands
p.a.vogt@uvt.nl

ABSTRACT

This paper explores the use of a social robot for one-on-one tutoring, in a study in which 15 children participated in four second-language tutoring sessions. Specifically, changes across sessions are measured on two dimensions: engagement and performance. Results have revealed a significant positive change in performance as well as a significant pattern in engagement across the interactions[1].

KEYWORDS

Child-robot interaction, language tutoring, education

ACM Reference format:

Emmy Rintjema, Rianne van den Berghe, Anne Kessels, Jan de Wit and Paul Vogt. 2018. A Robot Teaching Young Children a Second Language: The Effect of Multiple Interactions on Engagement and Performance. In *HRI'18 Companion: Conference on ACM/IEEE International Conference on Human-Robot Interaction, March 5-8, 2018, Chicago, IL, USA.* ACM, NY, NY, USA, 2 pages. DOI: https://doi.org/10.1145/3173386.3177059

1 INTRODUCTION

In recent years, increasing effort is made to design social robots as second language (L2) tutors [1-2]. The potential for robots to be effective tutors comes from various aspects, including their ability to tutor one-on-one [3], and to interact with children over multiple sessions for a longer period of time. However, most L2 learning studies thus far involving one-on-one tutoring lasted only one session [4], while those that have carried out long-term studies allowed one-on-many tutoring [2]. The current study explores the effects of a robot tutor over multiple child-robot L2 tutoring sessions. Specifically, this study investigates potential changes in two dimensions: children's engagement and their performance during the tutoring sessions.

2 OUR APPROACH

This study has been carried out as an extensive pilot study as a part of the L2TOR project, which aims to develop a robot tutor that helps young children learn an L2 [5]. To explore the changing relation between child and robot over the course of multiple one-on-one tutoring sessions, we conducted an experiment that consists of four sessions. During these sessions, the robot taught English (L2) vocabulary to Dutch (L1) children (age 5-6). Specifically, we examined performance and engagement. Performance has been measured as the degree to which children managed to complete tasks during the sessions and a word-knowledge task during a post-test. Task engagement has been measured as the degree to which children are actively involved in the tutoring session. This way, the current study explores how performance and engagement change over time. Based on findings by previous studies [4-7], it is expected that engagement will decrease over the course of the sessions. However, children's performance on the tasks will not be affected, as children will likely become more relaxed and familiarized to the robot and the task setting over time.

3 METHODS

The participants were 15 Dutch children (5 girls and 10 boys) with an average age of 5 years and 6 months (SD = 4.6 months). All parents gave their consent. The experimental setup contained a SoftBank Robotics NAO robot, which interacted autonomously with the child in four L2 tutoring sessions. Prior to the sessions, an introduction session was organized at the school to introduce the robot to the children. During the first three sessions, the robot taught the children a total of 17 words related to math such as numbers, adjectives (e.g., big) and verbs (e.g., add), via several games that the child and the robot played together on a Microsoft Surface Pro 4. The fourth session was a recap lesson to repeat and consolidate all 17 target words. In all sessions, the robot had the role of a peer-tutor, that is, the child and the robot

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3177059

together played games on the tablet and 'learned' the words. At the end of each session (except for the recap lesson), the child had to complete a task. During this task, several items appeared on the tablet and the robot asked the child to tap on a specific item on the screen to which it referred in L2. The robot automatically logged children's answers to measure their performance on the tasks. In addition, to measure overall improvement of the child's English skill, a pre-test and two post-tests were administered. The pre-test was administered right before the first interaction. One post-test was completed immediately after the recap lesson and the other a week later. In these tests, we asked the children to translate the 17 target words from English to Dutch. The experiments were recorded using a camera. The changes in performance over the course of the sessions were measured by comparing the scores on the tasks that were completed at the end of each of the three learning sessions (the recap lesson did not contain a task part and was not included in this measure). In addition, the scores on the pre- and post-tests were compared to assess the child's knowledge of the target words. To measure changes in engagement over the course of the sessions, a perception study was conducted. Eleven participants (Mean age = 25, *SD* = 2.8) rated the task-engagement of children on a five-point differential scale for a total of 117 short video clips (5 seconds) without audio (2 per child per lesson, 3 videos were missing due to technical difficulties). A high degree of interrater reliability was found between the participants of the perception study. The average measured ICC was .886 ($F(116, 1160) = 8.74$, $p < .001$). The fragments were taken at specific moments in the robot's script, the first a few minutes after the start and the second a few minutes before the end of the lesson.

3 RESULTS

On average, scores on the immediate post-test ($M = 3.64$, $SD = 3.08$) were higher than scores on the pre-test ($M = 2.43$, $SD = 2.41$), but this difference was not significant ($Mdif = 1$, $t(14) = 1.46$, $p = .165$). The scores on the delayed post-test ($M = 4.21$, $SD = 3.14$) were significantly higher than the scores on the pre-test ($Mdif = 1.79$, $t(13) = 2.29$, $p = .039$), indicating that children's knowledge of the target words improved over time (max. possible score: 17). To test the relation between the amount of sessions with the robot and performance on the tasks and engagement, a one-way repeated measures ANOVA was performed for both dependent variables. For performance, the overall ANOVA was significant ($F(1,14) = 22.65$, $p < .001$, $\eta 2 = .72$), revealing a significant effect of session on performance. Performance increased between the first ($M = 0.64$, $SD = 0.35$) and second lesson ($M = 1.40$, $SD = 0.46$) ($Mdif = 0.96$, $p < .001$), but not between lesson two and lesson three ($M = 1.39$, $SD = 0.24$) (max. possible score: 2). For engagement, the overall ANOVA was also significant ($F(1,14) = 4.61$, $p = .014$, $\eta 2 = .02$), revealing a significant effect of session on engagement. No significant change was observed between the first ($M = 3.55$, $SD = 0.46$) and second lesson ($M = 3.64$, $SD = 0.36$). Engagement decreased between the second and third lesson ($M = 3.09$, $SD = 0.63$) ($Mdif = -0.55$, $p = .034$), and increased between the third and fourth lesson ($M = 3.62$, $SD = 0.62$) ($Mdif = 0.53$, $p = .039$) (max. possible score: 5).

4 CONCLUSIONS

In this study, we carried out an extensive pilot to study long-term effects of L2 tutoring using a social robot. Results have revealed a positive relationship between time spent with the robot and performance on the learning tasks. Children improved their learning achievements after spending more time with the robot, possibly because they get more used the robot as a tutor. However, we cannot know this for sure as the content of the individual lessons might have influenced the performance. Furthermore, results showed a decrease of engagement between the second and third session and an increase of engagement between the third and fourth session. The downward trend may be explained by familiarization with the robot. The positive change between the third and fourth session may have been caused by the content of the recap session being different from the three learning sessions or by the fact that children knew that it was the last time they got to play with the robot. While more extensive studies on changes in performance and engagement in longitudinal one-on-one tutoring need to be conducted, this explorative study has thus taken first steps and found promising results regarding how performance and engagement evolve over time in long-term child-robot L2 tutoring sessions. Moreover, this study has set the stage for a larger evaluation study planned in the near future involving more lessons, more children and different conditions.

Acknowledgements

We are grateful for the support of all L2TOR project members. In addition, we thank the school, parents and children for their participation. The research was financially supported by the EU H2020 L2TOR project (grant 688014).

REFERENCES

[1] Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martines, M., Das, M., Breazeal, C. (2016). Affective Personalization of a Social Robot Tutor for Chilren's Second Language Skills. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), 3951 – 3957.

[2] Kanda, T., Hirano, T., Eaton, D., Ishiguro, H.. (2004). Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. Human-Computer Interaction, 19, 61-84.

[3] VanLehn, K. (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4):197-221.

[4] De Wit, J., Schodde, T., Willemsen, B., Bergmann, K., De Haas, M., Kopp, S., Krahmer, E., Vogt. P. (2018). The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies. In Proceedings of ACM / IEEE International Conference on Human Robot Interaction, Chicago, Illinois USA, March 2018 (HRI 2018).

[5] Belpaeme, T. et al. (2015) L2TOR - Second Language Tutoring using Social Robots. In Proceedings of 1st Int. Workshop on Educational Robots. Springer.

[6] Leite, I., Martinho, C., & Paiva, A. (2013). Social Robots for Long-term Interaction: A Survey. International Journal of Social Robotics, 5(2), pp 291-308. DOI: https://doi.org/10.1007/s12369-013-0178-y

[7] Nalin, M., Baroni, I., Kruijff-Korbayova, I., Canamero, L., Lewis, M., Beck, A., Cuayahuitl, H., Sanna, A. (2012). Children's adaption in multi-session interaction with a humanoid robot. RO-MAN, 2012 IEEE, DOI: 10.1109/ROMAN.2012.6343778

Investigating the Effects of a Robot Peer on L2 Word Learning

Conference Paper · March 2018

DOI: 10.1145/3173386.3176986

6 authors, including:

Rianne van den Berghe
Utrecht University
5 PUBLICATIONS **1** CITATION
SEE PROFILE

Josje Verhagen
Utrecht University
37 PUBLICATIONS **220** CITATIONS
SEE PROFILE

Ora Oudgenoeg-Paz
Utrecht University
13 PUBLICATIONS **63** CITATIONS
SEE PROFILE

Investigating the Effects of a Robot Peer on L2 Word Learning

Rianne van den Berghe
Utrecht University
P.O. Box 80140, 3508TC Utrecht
The Netherlands
m.a.j.vandenberghe@uu.nl

Sanne van der Ven
Utrecht University
P.O. Box 80140, 3508TC Utrecht
The Netherlands
s.vanderven@uu.nl

Josje Verhagen
Utrecht University
P.O. Box 80140, 3508TC Utrecht
The Netherlands
j.verhagen@uu.nl

Ora Oudgenoeg-Paz
Utrecht University
P.O. Box 80140, 3508TC Utrecht
The Netherlands
o.oudgenoeg@uu.nl

Fotios Papadopoulos
Plymouth University
PL4 8AA, Plymouth
U.K.
fotios.papadopoulos@plymouth.ac.uk

Paul Leseman
Utrecht University
P.O. Box 80140, 3508TC Utrecht
The Netherlands
p.p.m.leseman@uu.nl

ABSTRACT

Previous research has shown that the presence of a human peer during a learning task can positively affect learning outcomes. The current study aims to find out how second language (L2) vocabulary gains differ depending on whether children are learning by themselves, with a child peer, or with a robot peer. Children were administered an L2 vocabulary training in one of these three conditions. Children's word learning was measured directly after the training and one week later. Contrary to our expectations, children learning by themselves outperformed children in the peer conditions on one out of four word knowledge tasks. On the other tasks, there were no differences between the three conditions. Suggestions to further study the potential benefits of a robot peer are provided.

KEYWORDS

Child-robot interaction; peer learning; L2 vocabulary learning

ACM Reference format:

R. van den Berghe, S. van der Ven, J. Verhagen, O. Oudgenoeg-Paz, F. Papadopoulos, and P. Leseman. 2018. Investigating the Effects of a Robot Peer on L2 Word Learning. In *HRI'18 Companion: Conference on ACM/IEEE International Conference on Human-Robot Interaction, March 5-8, 2018, Chicago, IL, USA.* ACM, NY, NY, USA, 2 pages.
DOI: https://doi.org/10.1145/3173386.3176986

1 INTRODUCTION

Human peers can positively affect learning outcomes, by transferring their knowledge onto the learner, increasing task enjoyment, or allowing for learning-by-teaching [1]–[3]. One of the advantages of robots over other forms of technology is that they can take up various roles in learning interactions, such as tutors, teaching assistants, and, crucially, peers. Perhaps robot can, similarly to human peers, enhance learning outcomes.

Present evidence in robot-assisted language learning studies on the effectiveness of robot peers is contradictory. Some studies employing a robot as a peer find that children do learn [4]–[6], while other studies find limited learning or effects for only a subgroup of the children (e.g., those who voluntarily continued playing with a robot over time) [7], [8]. In these studies, the presence of a robot peer has not always been systematically compared to children learning alone or together with a human peer, and they differ in their design (e.g., single or multiple sessions, the robot acting like a tutor versus a learner). The current study compares L2 vocabulary gains across three learning conditions: children learning by themselves, with a child peer, or with a robot peer during a single session. The findings will help develop effective robot peers.

2 METHOD

In this study, 67 Dutch kindergartners (26 girls and 41 boys) with an average age of 67 months (*SD* = 7) participated in an L2 (English) vocabulary training. They were randomly assigned to one of the three conditions: (1) the child-only condition, in which they were learning by themselves (*N*=23), (2) the robot-peer condition, in which they learned together with a robot (*N*=23), or (3) the child-peer condition, in which they learned together with a child of the same age (*N*=21).

Children were taught six L2 English target words: "heavy", "light", "full", "empty", "in front of", and "behind". As part of the training children had to manipulate 3D images of objects on a tablet (e.g., putting animals in a cage). In the child-only condition, the child performed all manipulations on the tablet screen. In the peer conditions, the target child and the robot or child peer took turns in performing actions on the tablet.

The robot used in the present study was a NAO robot, developed by Softbank Robotics. We used the Wizard-of-Oz approach. The robot's responses had been preprogrammed, such that its responses and behaviors were consistent for all children.

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
DOI: https://doi.org/10.1145/3173386.3176986

To make sure the children would perceive the robot as a peer, children were instructed prior to the training that the robot also did not know the English words yet and was going to learn these as well. The robot's behaviors were: 1) manipulating the tablet; 2) repeating the target words; 3) commenting on the children's manipulations; 4) pointing to the tablet while explaining what to do, in case the child failed a task. Types 1 and 2 were the same activities as the child (and child peer) was/were asked to do, type 3 was included to increase children's motivation and stimulate interaction, and type 4 was used to provide scaffolding.

Children's word learning gains were assessed immediately after the training and one week later using four tasks: (1) a translation task in which the child had to translate the word from English to Dutch; 2) the same task from Dutch to English; 3) a comprehension task in which children had to select the picture that best represented the target word out of four options; and 4) a sorting task in which the child had to sort cards depicting one of two antonyms, e.g. "heavy" and "light", into trays depending on the word depicted on it. In addition, children were asked directly after the training whether they perceived the robot as a friend or a teacher, to assess whether our framing of the robot as a peer succeeded. Finally, a non-word repetition task in which children repeated non-existing words [9] was administered during the delayed post-test to investigate the comparability of the three groups on an important skill related to word learning: phonological memory.

The training and the tests were administered individually in a quiet room at children's schools. The first session, in which the training and the immediate post-test were administered, lasted about 50 minutes. The second session, with the delayed post-test and the non-word repetition task, lasted about 30 minutes.

In our data analyses, we included phonological memory and age as covariates, as one-way ANOVAs revealed a significant difference across the three groups in phonological memory, p = .039, with post-hoc tests showing that children in the robot-peer condition outperformed children learning alone, and in age, p = .047, with post-hoc tests showing that children learning alone were older than children in the robot-peer condition. Due to floor effects, the scores on the translation tasks were transformed into a dichotomous variable (having produced no words or at least one word correctly).

3 RESULTS

First, we assessed whether framing the robot as a peer succeeded and whether children performed above chance level on the comprehension task (25%) and the sorting task (50%). Most children (18 out of 23) saw the robot as a friend rather than as a teacher. Children performed above chance level on the comprehension task and the sorting task in both sessions (all ps <.005, 1.04 < d < 2.03).

Pearson's Chi-Square Tests indicated no effect of condition on the scores of the translation tasks in both sessions (all ps > .101, .305 < φ < .382). A repeated-measures ANCOVA showed an interaction effect between condition and time (p = .012, partial η^2 = .10), which was significant for the comprehension task (p = .010, partial η^2 = .15), but not for the sorting task (p = .091, partial η^2 = .08). For the comprehension task, children learning alone outperformed children in the child-peer condition during the delayed post-test (p = .031, d = 0.72) (with a trend for children in the robot-peer condition, p = .071, d = 0.47), while they did not differ significantly from children in the peer conditions during the immediate test (both ps >.999, 0.12 < d < 0.14).

4 CONCLUSIONS

Contrary to our expectations, we found that children learning by themselves in an L2 vocabulary training outperformed children learning with a child or robot peer on one out of four word-knowledge tasks. On the other tasks, there were no differences between the three conditions.

A possible explanation for the lack of peer benefits is that the vocabulary training did not allow for enough interaction between the learner and the peer for the learner to benefit from the peer. In addition, there were fewer learning opportunities in both peer conditions by manipulating the tablet, as tasks were divided between the target child and the (child or robot) peer. We recommend future researchers to look into more interactive learning tasks in which robots can take a more active role in supporting children's learning. Furthermore, qualitative analyses, which were beyond the scope of the current paper, would be especially valuable to assess which types of interactional patterns in child-child and child-robot dyads do or do not benefit learning in such tasks.

ACKNOWLEDGMENTS

This study was carried out within the L2TOR project, funded by the European Union's H2020 program (grant no. 688014). We would like to thank the children, their parents, and the schools for their participation, and Lisa Limpens, Ilse Almekinders, Ellis Bouter, and Veerle Kalle for their help in collecting the data.

REFERENCES

[1] A. King, "Transactive peer tutoring: Distributing cognition and metacognition," *Educ. Psychol. Rev.*, vol. 10, no. 1, pp. 57–74, 1998.

[2] F. Yarrow and K. J. Topping, "Collaborative writing: The effects of metacognitive prompting and structured peer," *Br. J. Educ. Psychol.*, vol. 71, pp. 261–282, 2001.

[3] K. Topping, S. Hill, A. McKaig, C. Rogers, N. Rushi, and D. Young, "Paired reciprocal peer tutoring in undergraduate economics," *Innov. Educ. Train. Int.*, vol. 34, no. 2, pp. 96–113, 1997.

[4] E. Mazzoni and M. Benvenuti, "A robot-partner for preschool children learning English using socio- cognitive conflict," *Educ. Technol. Soc.*, vol. 18, no. 4, pp. 474–485, 2015.

[5] F. Tanaka and S. Matsuzoe, "Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning," *J. Human-Robot Interact.*, vol. 1, no. 1, pp. 78–95, 2012.

[6] S. Meiirbekov, K. Balkibekov, Z. Jalankuzov, and A. Sandygulova, "'You win, I lose': Towards adapting robot's teaching strategy," in *The Eleventh ACM/IEEE International Conference on Human-Robot Interaction*, 2016, pp. 475–476.

[7] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Hum. – Comput. Interact.*, vol. 19, no. 1, pp. 61–84, 2004.

[8] G. Gordon *et al.*, "Affective personalization of a social robot tutor for children's second language skills," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 3951–3957.

[9] J. Verhagen, E. De Bree, H. Mulder, and P. Leseman, "Effects of vocabulary and phonotactic probability on 2-year-olds' nonword repetition," *J. Psycholinguist. Res.*, vol. 46, pp. 507–524, 2017.

The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies

Jan de Wit
TiCC*
Tilburg University
j.m.s.dewit@uvt.nl

Thorsten Schodde
Faculty of Technology, CITEC‖
Bielefeld University
tschodde@techfak.uni-bielefeld.de

Bram Willemsen
TiCC*
Tilburg University
b.willemsen@uvt.nl

Kirsten Bergmann
Faculty of Technology, CITEC‖
Bielefeld University
kirsten.bergmann@uni-bielefeld.de

Mirjam de Haas
TiCC*
Tilburg University
mirjam.dehaas@uvt.nl

Stefan Kopp
Faculty of Technology, CITEC‖
Bielefeld University
skopp@techfak.uni-bielefeld.de

Emiel Krahmer
TiCC*
Tilburg University
e.j.krahmer@uvt.nl

Paul Vogt
TiCC*
Tilburg University
p.a.vogt@uvt.nl

ABSTRACT

This paper presents a study in which children, four to six years old, were taught words in a second language by a robot tutor. The goal is to evaluate two ways for a robot to provide scaffolding for students: the use of iconic gestures, combined with adaptively choosing the next learning task based on the child's past performance. The results show a positive effect on long-term memorization of novel words, and an overall higher level of engagement during the learning activities when gestures are used. The adaptive tutoring strategy reduces the extent to which the level of engagement is diminishing during the later part of the interaction.

KEYWORDS

Language tutoring; Robotics; Education; Human-Robot Interaction; Bayesian Knowledge Tracing; Non-verbal communication

ACM Reference Format:
Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2018. The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies. In *HRI '18: 2018 ACM/IEEE International Conference on Human-Robot Interaction, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3171221.3171277

1 INTRODUCTION

Robots show great potential in the field of education [24]. Embodied agents in the form of humanoid robots, in particular, may deliver educational content for various subjects in ways similar to human tutors. The main advantage of using such a robot compared to traditional learning tools is its physical presence in the referential world of the learner [20]. The human-like appearance and presence in the physical environment may facilitate interactions that are, to some extent, similar to the ways in which human teachers would communicate with their students. Care should be taken, however, to design for the correct amount of social behavior, so as to avoid distracting students from the task at hand [16].

When designing such interactions, we can draw upon ways in which human teachers give contingent support to students in their learning activities. For instance, particularly in one-on-one tutoring situations, teachers tend to adjust the pace and difficulty of learning tasks based on the past development and current skill set of the student [29]. For example, teachers may help by scaffolding, taking the initial knowledge base as a starting point and trying to optimize the learning gain by choosing the hardest task to perform that still lies within the zone of proximal development [32] of the student.

The use of gestures that coincide with speech is another way for teachers to provide scaffolding, particularly when the concepts which the gestures refer to are not yet mastered by the student [1]. For instance, when teaching a second language (L2), gestures can help to ground an unknown word in the target language by linking it iconically or indexically to a real world concept. Such a facilitating effect on word learning has been found for imitating gestures of a virtual avatar [2]. However, it is an open question if the embodied presence of a robot can be exploited to support language learning through a robot's gesturing, and if so, what kind of gestures would have a positive impact.

In this paper, we present the results of an experiment conducted to explore how these two tools for scaffolding the learning of language — choosing the task that yields the greatest potential learning gain for a particular student and the use of appropriate co-speech gestures — carry over to a humanoid robot. Both were combined

HRI '18, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-4953-6/18/03…$15.00
https://doi.org/10.1145/3171221.3171277

*Tilburg center for Cognition and Communication
‖Cluster of Excellence Cognitive Interaction Technology

in one study to better estimate what the relative importance of the respective techniques is, while keeping all other factors constant, and to find out whether the benefits of the two strategies can potentially reinforce or impede each other. The techniques were implemented and tested in a one-on-one tutoring system where children, four to six years old, play a game with a robot to learn an L2. In the next section, we briefly present the approaches taken to realize the adaptive tutoring along with co-speech gesturing of the robot. We then describe the experimental methodology, before reporting and discussing the results obtained.

2 BACKGROUND

2.1 Adaptive Bayesian Knowledge Tracing

A robot tutor that personalizes the learning experience for individual students has been shown to have a positive effect on performance [19]. This robot is also perceived as smarter or more intelligent and less distracting or annoying. In order to simulate the way human tutors tailor learning activities and difficulty levels to a particular student, an adaptive tutoring system would have to measure and track the knowledge level of the student. Often the knowledge is traced skill-wise, where in the case of language learning, the mastery of particular words or phrases in the target language is represented probabilistically (e.g., [11]). This approach yields promising results, but it lacks flexibility because of the need to define domain-specific distance metrics to choose the next skill. Others have used Dynamic Bayesian Networks to represent the learner's knowledge about a skill, conditioned on the past interaction and taking into account skill interdependencies [14]. This approach requires detailed knowledge about the learning domain to model those interdependencies and their parameters. Recently, Spaulding et al. [27] used a simpler approach based on Bayesian Knowledge Tracing (BKT) [6]. The general BKT model consists of latent variables S^t representing the extent to which the system believes a particular skill to be mastered by the student. The belief state of the system is updated based on observed variables O^t, which correspond to the result of a learning action (e.g., correctly or incorrectly answering a question), while accounting for possible cases of guessing $p(guess)$ and slipping $p(slip)$ during the answer process. It was shown that this model outperforms traditional approaches for tracing the knowledge state in learning interactions, and that it can be easily extended to, for example, incorporate the emotional state of a child. In previous work [26], we have extended the basic BKT with action nodes to also model the tutor's decision-making based on current beliefs about the student's knowledge state (see Figure 1). Additionally, we employed a latent variable S that can attain discrete values for each skill, corresponding to six bins for the belief state (0%, 20%, 40%, 60%, 80%, 100%). This allows for quantifying the robot's uncertainty about a learner's skills as well as the impact of tutoring actions on future observations and skills.

This so-called *Adaptive* Bayesian Knowledge Tracing (A-BKT) approach can be used to choose the next skill from which the learner will most likely benefit, by estimating the greatest expected knowledge gains. It tries to maximize the belief of each skill while also balancing over all skills and not teaching a particular skill over and over again, even if the answer to the task was wrong and the skill belief is the lowest. The system does not only allow to choose the best skill to address next, but also the action to be used for scaffolding the learning of this skill. In this context, actions can be, for example, different types of exercises, pedagogical acts, or task difficulties. For the sake of simplicity, three task difficulties have been established (easy, medium, hard) to address a skill and to find the best action for a given skill.

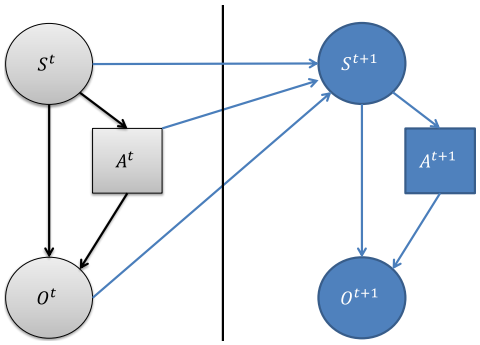


Figure 1: Dynamic Bayesian Network for BKT (taken from [26], with permission): with the current skill-belief the robot chooses the next skill S^t and action A^t for time step t and observes O^t as response from the user.

The goal of this strategy is to create a feeling of flow which can lead to better learning results [7]. It strives not to overburden the learner with tasks that would be too difficult nor to bore them with tasks that would be too easy, both of which may lead to disengagement and thus hamper the learning. Note that this approach is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system [18]. The implementation of A-BKT used in the current study is identical to the one used previously in [26]. However, it has not yet been evaluated with children nor in conjunction with other techniques that might affect action difficulty (such as gestures). Furthermore, its impact on student engagement has not been explored previously.

2.2 Gestures

Iconic gestures elicit a mental image that corresponds directly, either in form or execution, to the concept or action that is being described verbally at the same time [23]. For example, a flying bird could be depicted by stretching both arms sideways and moving them up and down. Studies have shown that iconic gestures, when performed by a human teacher, may aid the acquisition of L2 vocabularies [8, 15, 21, 28]. Hald et al. [12] provide an overview of how gestures can contribute to learning an L2. They propose that gestures might have a 'grounding' effect by linking existing perceptual and motor experiences to a new word. This is expected to result in a richer mental representation. Research by Rowe et al. [25] shows that gender, language background, and level of experience in the native language (L1) influence the extent to which gestures can contribute to L2 learning. The positive effects of gestures hold true for younger students as well; in fact, gestures are suggested to be a crucial part of communication with children [13]. It has also been

shown that gestures help not only to acquire knowledge, but also to retain it over time [5].

Previous research has explored the use of gestures by virtual agents (e.g., [2]) and robots (e.g., [30]), finding similar, positive effects on memory performance when gestures are produced by an artificial embodied agent compared to a human tutor. While humans tend to spontaneously perform and time their gestures, they will often need to be manually designed and coordinated with speech for the robot. Due to its limited degrees of freedom, however, the robot is unable to perform motions with the same level of detail, finesse, and accuracy as a human. This may lead to a loss in meaning when human gestures are being translated directly to the robot, indicating a need for alternative gestures. As a concrete example, the SoftBank Robotics NAO robot that was used in this case is unable to move its three fingers individually, preventing it from performing pointing gestures or finger-counting. However, research suggests that iconic gestures are almost as comprehensible when performed by a robot, compared to a human [4].

3 METHODOLOGY

An experiment was conducted to investigate the effect of using iconic gestures and an adaptive tutoring strategy on children's acquisition of L2 vocabularies, with the intention of answering the following three hypotheses:

> H1: There is a greater learning gain when target words are accompanied by iconic gestures during training, than in the case of not using gestures.
> H2: There is a reduced knowledge decay when target words are accompanied by iconic gestures during training, than in the case of not using gestures.
> H3: There is a greater learning gain when target words are presented in an adaptive order during training, based on the knowledge state of the child, than when target words are randomly introduced.

These hypotheses rely upon the underlying assumption that children are able to acquire new L2 words during a single session with a robot tutor, regardless of experimental conditions; this assumption was also put to the test.

The experiment had a 2 (adaptive versus non-adaptive) x 2 (gestures versus no gestures) between-subjects design. In the two conditions with the adaptive tutoring strategy, the A-BKT system described in Section 2.1 was used to select the target word for each round, based on the believed knowledge state of the child. In practice, this meant that children would be presented with a particular target word more frequently if they had answered it incorrectly in the past, thereby changing the number of times each target word occurred during training, although each target word was guaranteed to occur at least once. Other conditions had a random selection, where each of the six target words would always be presented five times, in a randomized order, for a total of thirty rounds. In the gesture conditions, whenever a target word was introduced in the L2 it was accompanied by an iconic gesture (as shown in Figure 2). All conditions had the robot standing up and in "breathing" mode, which meant that it slowly shifted its weight from one leg to the other and had a slight movement in its arms to simulate breathing.

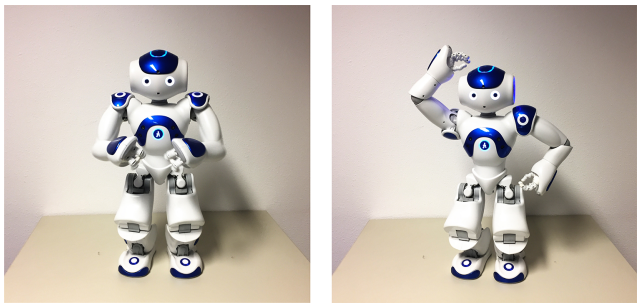


Figure 2: Examples of the stroke of two iconic gestures performed by the robot (taken from [9], with permission). Left: imitating a *chicken* by simulating the flapping of its wings; right: imitating a *monkey* by scratching head and armpit.

3.1 Participants

Participants were 61 children, with an average age of 5 years and 2 months ($SD = 7\,months$), 32 girls. They were recruited from primary schools in the Netherlands, by first contacting schools and then sending out an information letter together with a consent form through the schools to the parents of children that satisfied the age limit of four to six years. Only native Dutch children with Dutch as their L1 are included in the evaluation, although all 99 children that had signed up were allowed to participate in the experiment. The children were randomly assigned to conditions, while taking into account a balance in age and gender.

3.2 Materials

The aim of the tutoring interaction was to teach children six animal names in English: bird, chicken, hippo, horse, ladybug, and monkey. These specific words were chosen because the Dutch words are distinctly different from their English translations and because it was possible to create uniquely defining iconic gestures for them.

The SoftBank Robotics NAO robot was used, which was standing in front and slightly to the right of the child. After an experimenter had filled in the name of the child and pressed the start button, the experiment ran fully autonomously. Two experimenters were always present, where one would take care of getting the child from the classroom and explaining the procedure of the experiment, while the other would set up the system. To avoid having the child seek them out for feedback, the experimenters would announce that they would be occupied. The child was asked to sit on pillows, close to the tablet which was raised on a box and slightly tilted. Two cameras were used to record the interaction, one facing the front of the child and one at an angle from the side. The basic setup is shown in Figure 3, although it differed slightly between locations due to the layout of the rooms. In the condition with gestures every occurrence of the target word in L2, except when giving feedback, was accompanied by the matching iconic gesture (see Figure 2). The gesture was timed in such a way that the pronunciation of the target word would coincide with the stroke of the gesture, i.e., the accented phase that is most related to the meaning. A perception study was conducted to evaluate the quality of the gestures [9], where 14 participants were shown video recordings of all six gestures

Figure 3: The setup for the experiments.

performed by the robot and then asked to indicate which out of the six target words corresponds to each particular recording. Based on the results of this study, each gesture was deemed to be sufficiently unique to distinguish between the six target words.

The adaptive tutoring system starts with medium (0.5) confidence for all target words, a value associated with two distractors during training. Each distractor is a false answer to a task, an image belonging to one of the five other target words. In the random conditions, since there is no knowledge tracing the difficulty was always set to medium (two distractors). The tablet was used to get input from the child, because speech recognition does not work reliably with children [17]. This is also why only comprehension and not production of the target words is evaluated. An example of what the tablet screen would look like is shown in Figure 5. The images used during training belong to a different set of images than the ones used for the pre-test and post-tests. The set of images used during training matches the gesture that the robot performs related to the animals, for example the image of the horse for the training stage (shown in Figure 5) also includes a rider because the robot shows the act of riding a horse as a gesture. The image that was used during the tests did not include a rider and the horse is standing still, facing the opposite direction (shown in Figure 4). In addition to changing the pose or context of the animals, colors also varied. Together with having a recorded voice in the tests instead of the robot's synthesized speech, this aims to verify whether children learn how the English words map to the concepts of the animals and their matching Dutch words, rather than to one specific image.

3.3 Procedure

Prior to partaking in the experiment, participants were introduced to the robot during a group introduction. This approach is inspired by the work of Vogt et al. [31] with the intention of lowering the anxiety of children in subsequent one-on-one interactions with the robot. The introduction consisted of a description of what the robot is like, including a background story and how it is similar to humans in some respects, and different in others. Together with the children (and sometimes teachers and experimenters) the robot performed dances, after which all children were presented with

Figure 4: The pre-test and post-tests on a laptop, using a recorded voice and a different set of images from those on the tablet.

Figure 5: The tablet during training, showing images corresponding to the target word and two distractors.

the opportunity to shake the robot's hand before putting it to bed. Introductory sessions were scheduled several days before the first participant was to take part in the experiment, allowing time for the children to process these new impressions.

Before starting the tutoring interaction, a pre-test was administered to gauge the level of prior knowledge with respect to the animal names in the L1 (Dutch) and L2 (English). This test was administered on a laptop, where images of all six animals were randomly positioned on the screen. A recording of a (bilingual) native speaker pronouncing one of the six animal names was played, after which the child was asked to click the corresponding image on the screen (Figure 4). This was done for all six target words, first in Dutch and then in English.

After completing the pre-tests, the child would go through each target word one by one, still using the laptop. This is done to give the children a first exposure to the correct mappings between target words and the concepts they refer to, to avoid turning the first rounds of learning with the robot into a guessing game. Because there is no feedback during the pre-tests, this also ensures that concepts are linked to the correct word, rather than having the child assume that their answers during the pre-tests were all correct. For each word, the image of the corresponding animal would be shown in the center of the screen and the laptop would play a recording by a (bilingual) native speaker saying: "Look, this is a [target in L2]. Do you see the [target in L2]? Click on the [target in L2]!"

The training stage of the experiment consisted of the child and robot playing thirty rounds of the game *I spy with my little eye*. The robot, acting as the spy, would pick one of six target words and call out: "I spy with my little eye...", followed by the chosen word in the L2. For this stage, children were assigned to one of four conditions:

(1) Random tutoring strategy, no gestures ($N = 16$)
(2) Random tutoring strategy, gestures ($N = 14$)
(3) Adaptive tutoring strategy, no gestures ($N = 15$)
(4) Adaptive tutoring strategy, gestures ($N = 16$)

Prior to playing the game, the robot explained the procedure and asked the child to indicate whether they understood by pressing either a green or a red smiley. If the red smiley is pressed, the interaction would pause and an experimenter would step in to provide any further explanations. After this introduction, there were two practice rounds: one in Dutch and one in English.

After the robot had "spied" an animal, a corresponding image was shown on the tablet along with a number of distractor images (Figure 5). The child was then asked to pick the image that matched the animal name that the robot had spied. The number of distractors was determined by the difficulty level of the round, which in the case of the adaptive conditions depended on the confidence that the system had in that the child knew this particular target word. A low confidence resulted in only one distractor, while a high confidence had three distractors.

Feedback to the task was given by both the tablet and the robot. The tablet highlighted the image selected by the participant, either with a green, happy smiley if the correct answer was provided or a red, sad smiley if the selected image was an incorrect answer. The robot then provided verbal feedback, which in the case of a correct answer consisted of a random pick out of six positive feedback phrases (e.g., "well done!"), followed by "The English word for [target in L1] is [target in L2]". In the case of negative feedback, the robot would say "That was a [chosen answer in L1], but I saw a [target in L2]. [Target in L2] is the English word for [target in L1]". Whenever an incorrect answer was given, the same round would be presented once more but at the easiest difficulty (with only one distractor: the image that was incorrectly chosen in the previous attempt). This, combined with additional exposures in the corrective feedback, means that the number of times each target word was presented in the L2 may vary between children, depending on how many rounds were answered incorrectly. After finishing thirty rounds of training with the robot, the child was asked to complete a post-test on the laptop. This test is identical to the pre-test that was administered at the start of the experiment, in L2. Finally, the post-test was repeated once more, at least one week after the experiment, to measure long-term retention of the newly acquired knowledge.

3.4 Analysis

Immediate learning gain was measured as the difference between the number of correct answers on the post-test, administered directly after the training stage, and the number of correct answers on the pre-test, taken prior to the tutoring interaction. Test scores were always between 0 and 6 because each target word was asked once in the L2. The post-test was administered once more, (at least) one week after the experiment. We then looked at the difference between this delayed test and the pre-test for long-term learning gain. Finally, we took the difference between the delayed test and the immediate post-test as a measure of knowledge decay. The design of these tests is described in more detail in Section 3.2.

Children's tasks during training were of varying task difficulty in the adaptive tutoring condition, with one to three distractor images. To account for these differences, as well as to allow a comparison with the post-test results (five distractor images), we mapped binary task success (1: correct response; 0: incorrect response) onto the span between 0.0 and 1.0 by subtracting a value of 0.2 for each of the potential five distractor images that was not provided, which would, for example, result in a score of 0.6 for a correct response in a task with three distractors. The total score during training was then divided by the number of rounds (30), resulting in a training performance value between 0.0 and 1.0 (Figure 7).

4 RESULTS

The average duration of the training stage of the experiment was 18:38 minutes (SD = 3:03). Including the introduction, pre-test, and post-test this amounted to a session length of roughly thirty minutes. To confirm whether children managed to learn any new words from a single tutoring interaction, regardless of strategy or the use of gestures, a paired-samples t-test was conducted to measure the difference between post-test and pre-test scores for all conditions combined. There was a significant difference between the scores on the pre-test ($M = 1.75$, $SD = 1.14$) and immediate post-test ($M = 2.85, SD = 1.61$), $t(60) = 5.23, p < .001$. The same analysis was conducted for the delayed post-test that was taken (at least) one week after the experiment. Results revealed a significant difference between the pre-test scores ($M = 1.75, SD = 1.14$) and the delayed post-test test scores ($M = 3.02, SD = 1.40$), $t(60) = 6.81, p < .001$. However, there was no significant difference between the delayed post-test and the immediate post-test, $t(60) = .92, p = .34$. This means that H2 is not supported by these results, since no significant decay was observed in any of the conditions.

To investigate the effects of the different conditions on training performance, a two-way ANOVA was carried out with tutoring strategy (adaptive versus non-adaptive) and the use of gestures (gestures versus no gestures) as independent variables and performance during training as the dependent variable (Figure 7). As described in Section 3.4, these scores are weighted by the number of distractors present and divided by 30 rounds, resulting in a value between 0.0 and 1.0. For the 30 rounds of training there was a main effect of gesture use, $F(1, 57) = 18.23, p < .001, \eta_p^2 = .24$, such that training with gestures led to higher score ($M = .38, SD = .09$) than learning without gestures ($M = .29, SD = .08$). Children in the adaptive condition achieved a higher score ($M = .36, SD = .12$) than children in the non-adaptive condition ($M = .32, SD = .06$), but the effect of tutoring strategy was not significant, $F(1, 57) = 3.62, p = .06, \eta_p^2 = .06$. There was a significant interaction effect between use of gestures and tutoring strategy, $F(1, 57) = 4.72, p = .03, \eta_p^2 = .08$. Without gesture use, there was no significant difference between tutoring strategies. When gestures were present, however, children in the adaptive condition turned out to perform better than those in the non-adaptive condition. Hence, children's learning outcome was best when gesture use and adaptive training were combined.

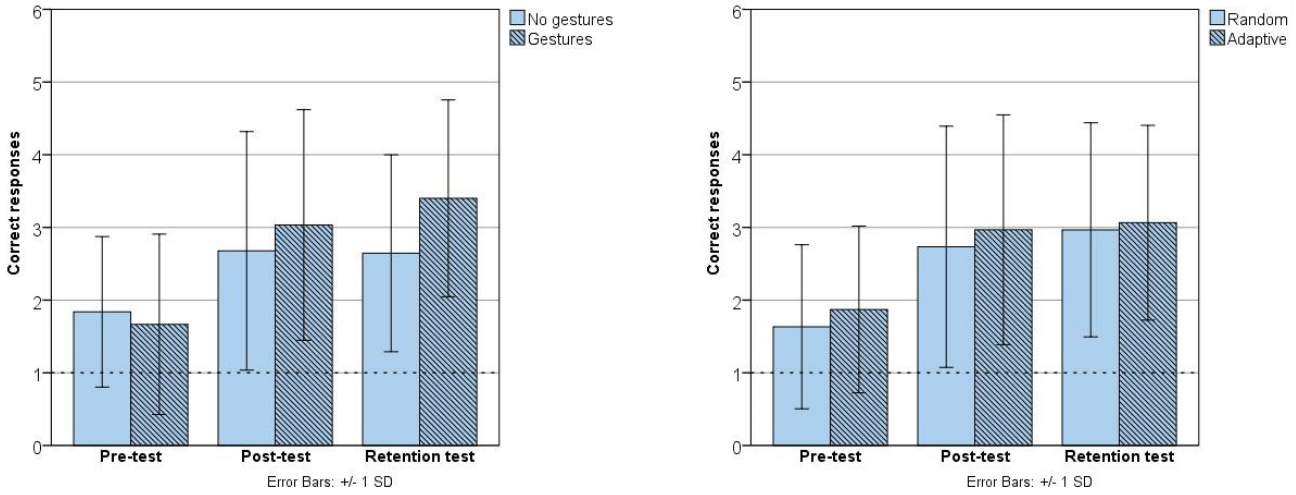


Figure 6: Test scores for the gesture vs no gesture conditions (left) and the adaptive vs random conditions (right).

Another two-way ANOVA was carried out to measure learning gain, with the difference score between the post-test results and the pre-test results as the dependent variable (Figure 6). There was no significant effect of tutoring strategy, $F(1, 57) < .001, p = .95, \eta_p^2 < .001$, or use of gestures, $F(1, 57) = 1.53, p = .22, \eta_p^2 = .03$. These results do not support H1 and H3 (greater learning gains when gestures and adaptive tutoring are used). The same two-way ANOVA with the difference score between results of the delayed post-test and the pre-test also did not give a significant effect of tutoring strategy, $F(1, 57) = .36, p = .55, \eta_p^2 = .006$, but there was a significant effect for use of gestures, $F(1, 57) = 6.11, p = .02, \eta_p^2 = .097$, indicating that the learning gain between pre-test and delayed post-test was greater when gestures were used during training ($M = 1.70, SD = 1.56$) than when no gestures were used ($M = .81, SD = 1.25$). Although this does not fully support H1 or H2, it does show a long-term learning gain when gestures are used during learning. No interaction effect was found, $F(1, 57) = .04, p = .84, \eta_p^2 \leq .001$.

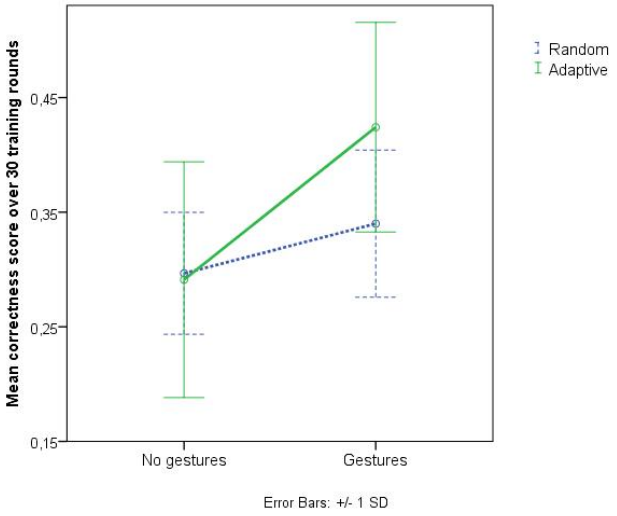


Figure 7: Interaction effects of gesture use and training strategy.

4.1 Evaluation of engagement

The engagement of the children during the training stage with the robot was examined to find out whether children became more disengaged with the tutoring tasks towards the end of the thirty rounds, and whether the application of an adaptive tutoring strategy and gestures would influence the change in engagement levels. This was done by asking 18 adult participants, without specific training in working with children, to rate video clips (without audio) of the children interacting with the robot. The choice for conducting a perception study with adults using video recordings of the experiment was made for two reasons: so that the training would not have to be interrupted for questions regarding the experience, thereby potentially influencing the engagement, and because it is difficult for children of a young age to reflect upon their experiences and verbalize these thoughts [22]. For each child, one clip was taken from the fifth round of training and one clip from the twenty-fifth round, to get observations that are close to the beginning and end of the training, but far enough from these actual moments to avoid short bursts of engagement when children realize the experiment is starting or finishing. The clips start right after the robot finishes introducing the task, i.e., the point at which the turn switches to the child to provide an answer. All clips then run for five seconds. One child that was excluded from the previous analysis because delayed post-test results were missing, was included for this part of the evaluation. However, data from one other child was missing, making the number of stimuli 122 (61 children, two clips each), with 14 to 16 children in each condition. Participants in the evaluation were asked to rate all 122 clips, randomly presented to them, on a scale from 1 (completely disengaged) to 7 (completely engaged). As a practice round, two clips of a child that was not included in the

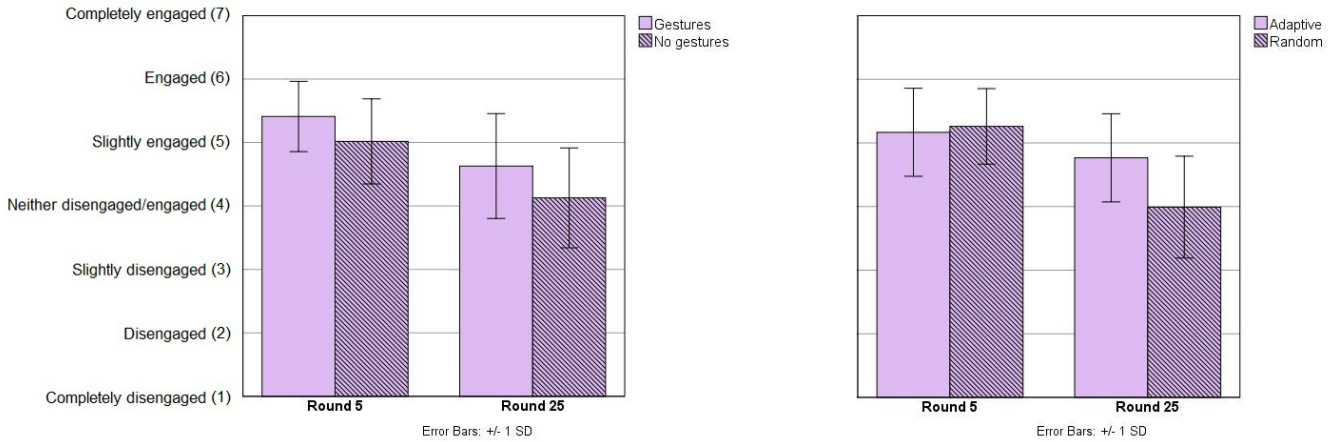


Figure 8: Rated engagement levels early and late in the training interaction for the gesture versus no gesture conditions (left) and the adaptive versus random conditions (right).

main experiment were presented, where one example was clearly engaged and the other was clearly not engaged. After this practice round, participants were told which features from the examples showed engagement (i.e., rapid response to the question, upright body posture, displaying joy after answering the question) and disengagement (i.e., slower response to the question, supporting the head by leaning on the arms, showing less interest in the task).

For each participant, the ratings were averaged over all children belonging to the same experimental condition, resulting in a total of eight average ratings (four conditions, each with fifth and twenty-fifth round). Figure 8 visualizes the data from the evaluation. Results from a paired-samples t-test showed that children were considered to be significantly less engaged in the twenty-fifth round ($M = 4.38, SD = .84$) than in the fifth round ($M = 5.21, SD = .64$), $t(71) = -12.09, p < .001$. Furthermore, a two-way ANOVA with tutoring strategy (adaptive versus non-adaptive) and gesture use (gestures versus no gestures) as factors showed no significant effect for the use of gestures, $F(1, 68) = 1.36, p = .25, \eta_p^2 = .02$, but there was a significant effect for tutoring strategy, $F(1, 68) = 86.26, p < .001, \eta_p^2 = .559$. The drop in engagement between round five and round twenty-five was less when an adaptive strategy was applied ($M = -.40, SD = .35$) than when words were randomly presented ($M = -1.27, SD = .44$). There was no interaction effect between gestures and tutoring strategies, $F(1, 68) = .01, p = .93, \eta_p^2 = .00$. The same analysis was conducted with the average engagement level of the fifth and twenty-fifth rounds combined, to get an idea of the overall engagement throughout the entire training session in different conditions. In this case the overall level of engagement was significantly higher in the gesture condition ($M = 5.02, SD = .63$) than in the condition without gestures ($M = 4.57, SD = .68$), $F(1, 68) = 8.75, p = .004, \eta_p^2 = .114$. There was also a significantly higher engagement when an adaptive strategy was used ($M = 4.97, SD = .67$) as opposed to a random tutoring strategy ($M = 4.63, SD = .67$), $F(1, 68) = 5.10, p = .03, \eta_p^2 = .07$. No interaction effect between the two factors was found, $F(1, 68) = .08, p = .78, \eta_p^2 = .001$.

5 DISCUSSION

The results presented above show that by spending a single tutoring interaction of about twenty minutes with a robot tutor, young children were able to acquire new words in an L2, regardless of the experimental condition, and were also able to retain this newly acquired knowledge for a prolonged period of time. Care was taken to design the pre-test and post-tests in such a way to be clearly distinct from the training session with the robot in terms of physical context (laptop versus tablet), voice, and characteristics of the images used, with the aim of getting a reliable measure of the attained knowledge. Results from the pre-test show that there is indeed a realistic amount of prior knowledge, on average above chance, presumably because some children have been exposed previously to the target words, for example in television programs. The observed number of correct answers on the immediate and delayed post-test are higher than on the pre-test, indicating the expected knowledge gain after engaging in learning activities. The scores on the post-test are lower than the number of correct answers towards the end of the training stage, which could show that indeed the test evaluates whether children acquire the underlying concepts, rather than simply being able to link a word being pronounced by the robot to one specific image (in some cases with the help of gestures that are not present in the tests). One potential point of improvement for the tests could be to introduce context when querying the target words, for example by using sentences rather than isolated words. Although explicitly instructed, children seemed not always aware that they were supposed to select the image corresponding to an *English* word, causing them to choose the animal with the most similar sounding name in Dutch instead (e.g., bird was often confused with the Dutch word 'paard').

When gestures were performed by the robot during training, there was a higher retention of newly acquired words after at least one week. This aligns with similar effects that were shown previously in the context of math with a human tutor [5] and indicates that these indeed carry over to a robot; a compelling finding that warrants future research into the intricacies of gesture use by humanoid robots. As mentioned by Hostetter [13] with respect to

human-human communication, it appears that gestures retain their positive effects on communication when they are scripted rather than being produced spontaneously. In this work, only iconic gestures are used that clearly relate to the concept they describe. Future work could investigate whether a similar contribution to learning gain is found when non-iconic gestures are used. Furthermore, the target words used in this experiment were chosen specifically such that matching gestures could be designed for the robot. It would be interesting to explore how well a broader range of gestures, describing various abstract and concrete concepts, could be performed by a robot as opposed to a human interlocutor. Finally, asking children to actually re-enact the gestures (e.g., as in [8, 28]), or to come up with their own gestures, might further increase the potential utility of gestures in learning due to the embodiment effect [10].

The test results regarding the adaptive tutoring system are currently inconclusive. This might be a result of the manner in which learning gain was measured, i.e., a quantification of newly acquired words — perhaps the adaptive system did not result in *more* words learned, but rather led to a more focused acquisition of exactly those words that the child found most difficult. The main remaining difference between the ways in which human teachers and the system presented here personalize content is that teachers tend to draw upon a memory that spans a longer period of time. In this experiment, the memory of the adaptive system was built up, and then applied, over the course of a single session. The system might come to fruition if there are multiple sessions with the same child, allowing the results of one session to become prior knowledge for the next one. It is also possible that the actions that the system performs based on the estimated knowledge levels of the child are too subtle. Currently, only the order and frequency of words is tailored, within the thirty rounds, and different levels of difficulty are represented by adding or removing one distractor image. Actions and difficulty levels could be more complex than that, for example by applying completely different tutoring strategies or games that might fit a particular child better. For the sake of this experiment, the number of rounds was fixed to thirty, but this session length might also be left up to the adaptive system to control. This would allow the interaction to end at the exact moment where the learning is 'optimal', i.e., a point at which the adaptive system thinks that the child has achieved his or her highest potential learning gain. A final avenue for improvement that is currently being pursued is to incorporate additional information about the affective state of the child. Some children might not be in the right mood to learn when they start, or their attention might fade during the interaction; rather than focusing only on the learning objectives the robot might want to engage in activities that work towards creating and maintaining the right atmosphere for learning.

We found it valuable to include the measure of children's engagement during the interaction. A higher level of engagement indicates increased motivation and willingness to learn [3]. Although students might succeed in simple word learning with limited engagement and the use of a low-level learning strategy, increased engagement could stimulate them to go beyond simple memorization and relate these new words to prior knowledge. Furthermore, engagement can serve as a measure of how well the learning activities are tailored to the child's abilities — constantly presenting tasks that are either too hard or too easy could have a detrimental effect on engagement. The results of our evaluation show that indeed the adaptive system appears to match the learning activities to each child's needs by providing a realistic yet challenging task, resulting in a reduced decline in engagement towards the end of the interaction. Gestures contribute to a higher overall engagement, which could be explained by the fact that the robot appears more active and playful in this condition, thereby stimulating the child to remain engaged.

6 CONCLUSION

The study presented in this paper aimed to explore if a humanoid robot can support children, four to six years old, in learning the vocabulary of a second language. We found that, indeed, children manage to learn new words during a single tutoring interaction, and are able to retain this knowledge over time. Specifically, we investigated whether the effects of tailoring learning tasks to the knowledge state of the learner and using co-speech gestures — both of which are strategies used by human teachers to scaffold learning — transfer to the use of a humanoid robot tutor. Our results show that the robot's use of gestures has a positive effect on long-term memorization of words in the L2, measured after one week. Furthermore, children appear more engaged throughout the tutoring session and are able to provide more correct answers when gestures are used. An adaptive tutoring strategy helps to reduce the drop in engagement that inevitably happens over the course of an interaction, by providing contingent, personalized support to each learner. By combining both methods in a tutoring session, adaptivity seems to succeed in finding the 'sweet spot' of challenging children enough to keep them motivated while gestures can add to overall engagement and support children in finding the correct answer. Therefore, gestures can form an additional tool in the toolbox of A-BKT to be deliberately employed, for example, when a reduced difficulty is deemed necessary or engagement is decreasing.

ACKNOWLEDGMENTS

This work is partially funded by the H2020 L2TOR project (grant 688014), the Tilburg center for Cognition and Communication 'TiCC' at Tilburg University (Netherlands) and the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), funded by the German Research Foundation (DFG), at Bielefeld University (Germany). The authors would like to thank all members of the L2TOR project for their valuable comments and suggestions that have contributed towards the design of the experiment. Furthermore, we are grateful to the schools, parents, and children that participated in our experiment, Elske van der Vaart for lending us her voice for the content on the laptop, as well as Sanne van Gulik, Marijn Peters Rit, and Emmy Rintjema for their help with data collection. The preliminary design of this experiment was first presented at the R4L workshop, HRI'17 [9]; we thank the attendees for their feedback.

REFERENCES

[1] Martha W. Alibali and Mitchell J. Nathan. 2007. Teachers' Gestures as a Means of Scaffolding Students' Understanding: Evidence From an Early Algebra Lesson. *Video Research in the Learning Sciences* 39, 5 (2007), 349–366. https://doi.org/10.1111/j.1467-8535.2008.00890_7.x

[2] Kirsten Bergmann and Manuela Macedonia. 2013. A virtual agent as vocabulary trainer: iconic gestures help to improve learnersâĂŹ memory performance. In *International Workshop on Intelligent Virtual Agents*. Springer, 139–148.

[3] Phyllis C. Blumenfeld, Toni M. Kempler, and Joseph S. Krajcik. 2005. *Motivation and Cognitive Engagement in Learning Environments*. Cambridge University Press, Cambridge, Chapter 28, 475–488. https://doi.org/10.1017/CBO9780511816833.029

[4] Paul Bremner and Ute Leonards. 2016. Iconic gestures for robot avatars, recognition and integration with speech. *Frontiers in Psychology* 7 (feb 2016), 183. https://doi.org/10.3389/fpsyg.2016.00183

[5] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. 2008. Gesturing makes learning last. *Cognition* 106, 2 (2008), 1047–1058. https://doi.org/10.1016/j.cognition.2007.04.010 arXiv:NIHMS150003

[6] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[7] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of educational media* 29, 3 (2004), 241–250.

[8] Jacqueline A. de Nooijer, Tamara van Gog, Fred Paas, and Rolf A. Zwaan. 2013. Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. *Acta Psychologica* 144, 1 (2013), 173–179. https://doi.org/10.1016/j.actpsy.2013.05.013

[9] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2017. Exploring the Effect of Gestures and Adaptive Tutoring on Children's Comprehension of L2 Vocabularies. In *Proceedings of the Workshop R4L at ACM/IEEE HRI 2017*.

[10] Katinka Dijkstra and Lysanne Post. 2015. Mechanisms of embodiment. 6, OCT (2015), 1525. https://doi.org/10.3389/fpsyg.2015.01525

[11] Goren Gordon and Cynthia Breazeal. 2015. Bayesian Active Learning-based Robot Tutor for Children's Word-reading Skills. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)*. AAAI Press, 1343–1349.

[12] Lea A. Hald, Jacqueline de Nooijer, Tamara van Gog, and Harold Bekkering. 2016. Optimizing Word Learning via Links to Perceptual and Motoric Experience. *Educational Psychology Review* 28, 3 (2016), 495–522. https://doi.org/10.1007/s10648-015-9334-2

[13] Autumn B. Hostetter. 2011. When do gestures communicate? A meta-analysis. *Psychological Bulletin* 137, 2 (2011), 297–315. https://doi.org/10.1037/a0022128

[14] Tanja Käser, Severin Klingler, Alexander Gerhard Schwing, and Markus Gross. 2014. Beyond knowledge tracing: Modeling skill topologies with bayesian networks. In *International Conference on Intelligent Tutoring Systems*. Springer, 188–198.

[15] Spencer D. Kelly, Tara McDevitt, and Megan Esch. 2009. Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes* 24, 2 (2009), 313–334. https://doi.org/10.1080/01690960802365567 arXiv:http://dx.doi.org/10.1080/01690960802365567

[16] James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 67–74. https://doi.org/10.1145/2696454.2696457

[17] James Kennedy, Severin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child Speech Recognition in Human-Robot Interaction : Evaluations and Recommendations. *Proc. of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2017), 82–90. https://doi.org/10.1145/2909824.3020229

[18] S. Leitner. 1972. *So lernt man Lernen: Der Weg zum Erfolg [Learning to learn: The road to success]*. Freiburg: Herder.

[19] Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. 2014. Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 423–430.

[20] Daniel Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. *34th Annual Conference of the Cognitive Science Society* 34, 1 (jan 2012), 1882–1887. https://doi.org/ISBN978-0-9768318-8-4

[21] Manuela Macedonia, Karsten Müller, and Angela D. Friederici. 2011. The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping* 32, 6 (2011), 982–998. https://doi.org/10.1002/hbm.21084

[22] Panos Markopoulos, Janet C. Read, Stuart MacFarlane, and Johanna Hoysniemi. 2008. *Evaluating Children's Interactive Products: Principles and Practices for Interaction Designers*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Chapter 1, 3–18.

[23] David McNeill. 1985. So you think gestures are nonverbal? *Psychological Review* 92, 3 (1985), 350–371. https://doi.org/10.1037/0033-295x.92.3.350

[24] Omar Mubin, Catherine J. Stevens, Suleman Shahid, Abdullah Al Mahmud, and Jian-Jie Dong. 2013. A Review of the Applicability of Robots in Education. *Technology for Education and Learing* 1 (2013), 209—-0015. https://doi.org/10.2316/Journal.209.2013.1.209-0015

[25] Meredith L. Rowe, Rebecca D. Silverman, and Bridget E. Mullan. 2013. The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemporary Educational Psychology* 38, 2 (2013), 109–117. https://doi.org/10.1016/j.cedpsych.2012.12.001

[26] Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. 2017. Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. In *Proceedings of ACM/IEEE HRI 2017*. ACM Press, 128–136. https://doi.org/10.1145/2909824.3020222

[27] Samuel Spaulding, Goren Gordon, and Cynthia Breazeal. 2016. Affect-Aware Student Models for Robot Tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems (AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 864–872.

[28] Marion Tellier. 2008. The effect of gestures on second language memorisation by young children. *Gesture* 8, 2 (2008), 219–235. https://doi.org/10.1075/gest.8.2.06tel

[29] Janneke van de Pol, Monique Volman, and Jos Beishuizen. 2010. Scaffolding in teacher-student interaction: A decade of research. (2010), 271–296 pages. https://doi.org/10.1007/s10648-010-9127-6 arXiv:arXiv:1002.2562v1

[30] Elisabeth T. van Dijk, Elena Torta, and Raymond H. Cuijpers. 2013. Effects of Eye Contact and Iconic Gestures on Message Retention in Human-Robot Interaction. *International Journal of Social Robotics* 5, 4 (2013), 491–501. https://doi.org/10.1007/s12369-013-0214-y

[31] Paul Vogt, Mirjam De Haas, Chiara De Jong, Peta Baxter, and Emiel Krahmer. 2017. Child-Robot Interactions for Second Language Tutoring to Preschool Children. *Frontiers in human neuroscience* 11, March (2017), 1–7. https://doi.org/10.3389/fnhum.2017.00073

[32] Lev Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.

International Journal of Social Robotics (2018) 10:325–341
https://doi.org/10.1007/s12369-018-0467-6

Guidelines for Designing Social Robots as Second Language Tutors

Tony Belpaeme[1,2] · Paul Vogt[3] · Rianne van den Berghe[5] · Kirsten Bergmann[4] · Tilbe Göksun[6] · Mirjam de Haas[3] · Junko Kanero[6] · James Kennedy[1] · Aylin C. Küntay[6] · Ora Oudgenoeg-Paz[5] · Fotios Papadopoulos[1] · Thorsten Schodde[4] · Josje Verhagen[5] · Christopher D. Wallbridge[1] · Bram Willemsen[3] · Jan de Wit[3] · Vasfiye Geçkin[6] · Laura Hoffmann[4] · Stefan Kopp[4] · Emiel Krahmer[3] · Ezgi Mamus[6] · Jean-Marc Montanier[7] · Cansu Oranç[6] · Amit Kumar Pandey[7]

Accepted: 11 January 2018 / Published online: 25 January 2018

Abstract
In recent years, it has been suggested that social robots have potential as tutors and educators for both children and adults. While robots have been shown to be effective in teaching knowledge and skill-based topics, we wish to explore how social robots can be used to tutor a second language to young children. As language learning relies on situated, grounded and social learning, in which interaction and repeated practice are central, social robots hold promise as educational tools for supporting second language learning. This paper surveys the developmental psychology of second language learning and suggests an agenda to study how core concepts of second language learning can be taught by a social robot. It suggests guidelines for designing robot tutors based on observations of second language learning in human–human scenarios, various technical aspects and early studies regarding the effectiveness of social robots as second language tutors.

Keywords Social robot · Second language learning · Robot tutor · Human–robot interaction

1 Introduction

One of the goals of Human–Robot Interaction (HRI) is to research and develop autonomous social robots as tutors that are able to support children learning new skills effectively through repeated interactions. To achieve this, the interactions between child and robot should be pleasant, challenging, and pedagogically sound. Interactions need to be pleasant for children to enjoy, challenging so that children remain motivated to learn new skills, and pedagogically sound to ensure that children receive input that optimises their learning gain. One domain in which robots for learning are developed is second language (L2) tutoring (e.g., [1,33,64]). While much progress has been made in this field, there has not been an effective one-on-one L2 tutoring programme that can be structurally applied in educational settings for various language communities.

The L2TOR project[1] (pronounced as 'el tutor') aims to bridge this gap by developing a lesson series that helps preschool children, around the age of 5 years, learn basic vocabulary in an L2 using an autonomous social robot as tutor [8]. In particular, we develop one-on-one, personalised interactions between children and the SoftBank NAO robot for teaching English to native speakers of Dutch, German, and Turkish, and for teaching Dutch or German to Turkish-speaking children living in the Netherlands or Germany. To ensure a pedagogically sound programme, lessons are being developed in close collaboration with developmental psychologists and pedagogists.

Paul Vogt
p.a.vogt@uvt.nl

1 Centre for Robotics and Neural Systems, Plymouth University, Plymouth, UK

2 IDLab – imec, Ghent University, Ghent, Belgium

3 Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands

4 Cluster of Excellence Cognitive Interaction Technology, Bielefeld University, Bielefeld, Germany

5 Department of Special Education: Cognitive and Motor Disabilities, Utrecht University, Utrecht, The Netherlands

6 Department of Psychology, College of Social Sciences and Humanities, Koç University, Istanbul, Turkey

7 SoftBank Robotics, Paris, France

[1] http://www.l2tor.eu.

Personalising the interactions between child and robot is crucial for successful tutoring [45]. Personalisation can be achieved by creating some initial common ground between child and robot, and by having the robot adapt to the individual progress of children. Constructing initial common ground helps to promote long-term interactions between child and robot [33], and can be achieved by framing the robot as a peer and by explaining (dis)similarities between robots and humans. However, to keep children motivated to learn, it is important to keep the learning targets within the child's *Zone of Proximal Development* [70]. Throughout the lessons the target should be sufficiently challenging for the child: not too challenging as this may frustrate the learner and not too easy as this may bore the learner. Moreover, interactions should be designed such that the robot provides a scaffold that allows the child to acquire the desired language skills. For instance, by providing non-verbal cues (e.g., gestures) that help to identify a word's referent or by providing appropriate feedback, it is possible for children to reinforce successfully acquired skills or to correct suboptimal (or wrong) skills.

The L2TOR approach relies on the current state-of-the-art in HRI technology, which offers promising opportunities, but also poses major challenges. For instance, NAO has the ability to produce speech in various languages, making it possible for the robot to address the child in both the native language (L1) and in the L2. However, at present, automatic speech recognition (ASR) for child speech is not performing to a sufficiently reliable standard, and thus using ASR is currently infeasible [37]. This not only limits the ability to rely on verbal interactions since the robot is unable to respond to children's speech, but it also limits the ability to monitor and respond to children's L2 productions. Hence, our design has to find ways to work around such technological limitations.

The paper aims to present a number of guidelines that help researchers and developers to design their own social robot, especially for, though not necessarily limited to, L2 tutoring. After a brief review of L2 learning from a developmental psychology point of view, Sect. 3 reviews some previous research on language tutoring using social robots. In Sect. 4, we will present our guidelines relating to pedagogical considerations, child–robot interactions and interaction management. These issues will be discussed in light of some of our early experiments. Section 5 discusses our approach to evaluating the L2TOR system, which is designed to demonstrate the (potential) added value of using social robots for L2 tutoring.

2 Second Language Learning

Learning an L2 is important in today's society. In the European Union (EU), for example, 54 percent of the population can hold a conversation in at least two languages, and 25 percent are able to speak three languages [20]. Consequently, L2 teaching has become an essential part of primary education. In 2002, the EU proposed a multilingualism policy of teaching an L2 to all young children. The policy suggests every European citizen learns practical skills in at least two languages aside from their L1 [4]. According to a recent survey, the vast majority of European citizens (98 percent of the respondents in this survey) believe that mastering a foreign language is useful for the future of their children [20].

Preschool years are vital for L2 learning, because later academic success depends on early language skills [29]. For children learning English as their school language, their English vocabulary size predicts their performance in English reading tests [57]. Although learning an L2 comes naturally for some children, for many others it is a challenge that they must overcome. For children from immigrant families or minority communities, the language used at school is often different from the language used at home. These children, thus, not only start learning the school language later than their peers, but also continue to receive relatively less input in each of their languages [30]. Hence, novel ways to expose children to targeted L2 input must be considered.

Patterns of L2 learning largely mirror those of L1 learning, which requires both the quantity and the quality of language input to be sufficient [27]. Children do not learn language just by listening to speech; rather, interactive experience is essential [39]. L2 learning is no exception, and several factors such as interactivity must be considered (see [38] for a review). In addition to quantity, socio-pragmatic forms of interaction involving joint attention, non-verbal interaction, feedback, and temporal and semantic contingencies are expected to contribute to L2 learning [3,9,59,66]. However, there are also some notable differences between L1 and L2 learning. For example, in L2 education it is important to consider from whom children are learning the L2. Place and Hoff [56] found that hearing English from different speakers and the amount of English input provided by native speakers is critical for learning English as L2. Another notable difference between L1 and L2 learning is that children may rely on their L1 when learning an L2 (e.g., [75]). Thus, we may need to be cautious about factors such as negative transfer or interference, in which some concepts and grammar in the L2 are hard to acquire because children are thinking in their L1 [67].

When children are learning more than one language, the amount of input a child hears in each language predicts vocabulary size in each language [30,55]. Bilingual children tend to have a smaller vocabulary size in each language compared to their monolingual peers [54], although the combined or conceptual vocabulary size of both languages is often equal to that of monolinguals [31,54]. The amount of language input also affects language processing speed and trajectories of vocabulary learning, and thus early language

input may have cascading effects on later language learning. Hurtado et al. [32] found that the amount of language input bilingual children receive at 18 months of age predicts their speed of recognizing words and the size of their vocabulary at 24 months. To properly foster development of two or more languages, adults must carefully consider a good balance between languages [67].

Although both monolingual and bilingual children monitor and respond to social pragmatic cues, bilingual children have heightened sensitivity to those non-linguistic cues, probably due to an early communicative challenge they face because of less than perfect mastery in one of the languages [74]. Brojde et al. [10] found that bilingual children rely more on eye gaze than their monolingual counterparts when learning novel words. Yow and Markman [76] also demonstrated that 3- and 4-year-old bilingual children were better at understanding and using gestures and gaze direction to infer referential intent. Thus, especially for children with advanced L2 knowledge, we may be able to boost their learning process by making use of these pragmatic cues.

As the demand for early L2 education increases, the usage of additional teaching opportunities in terms of educational tablet games, or electronic vocabulary trainers becomes more and more important to increase the quantity of L2 input. Moreover, especially with regard to young children, the consideration of embodied technologies (e.g., virtual agents or robots) seems reasonable, because they invite intuitive interactions that would add to the quality of the L2 input. The question then becomes: how should such a robot be designed?

3 Robots for Language Tutoring

In recent years, various projects have started to investigate how robot tutors can contribute to (second) language learning. In this section, we review some of these studies, focusing on: (a) the evidence that robots can promote learning; (b) the role of embodiment in robot tutoring; and (c) the role of social interactions in tutoring.

3.1 Learning from Robots

There has been an increased focus on how social robots may help engage children in learning activities. Robots have been shown to help increase interaction levels in larger classrooms, correlating with an improvement in children's language learning ability [22]. How best to apply this knowledge in the teaching of a foreign language has been explored by different researchers from various perspectives. Alemi et al. [1] employed a social robot as an assistant to a teacher over a 5-week period to teach English vocabulary to Iranian students. They found that the class with the robot assistant learned significantly more than that with just the human teacher. In addition, the robot-assisted group showed improved retention of the acquired vocabulary. This builds on earlier findings by [33] where a 2-week study with a robot situated in the classroom revealed a positive relation between interacting with a robot and vocabulary acquisition. Further results by [64] also confirm that the presence of a robot leads to a significant increase in acquired vocabulary. Movellan et al. [50] selected 10 words to be taught by a robot, which was left in the children's classroom for 12 days. At the end of the study, children showed a significant increase in the number of acquired words when taught by the robot. Lee et al. [42] further demonstrated that robot tutoring can lead not just to vocabulary gains, but also improved speaking ability. In their study, children would start with a lesson delivered by a computer, then proceed to pronunciation training with a robot. The robot would detect words with an expanded lexicon based on commonly confused phonemes and correct the child's pronunciation. Additionally, the children's confidence in learning English was improved.

All of these studies show the capacity of various robots as tutors for children (with the children's age ranging from 3 to 12 years old) learning an L1 or L2 'in the wild'. However, what exactly is it that gives robots the capacity for tutoring? Moreover, how does this compare to other digital technologies, such as tablets and on-screen agents? Is it merely the embodiment of the robot, or rather the quality of social interactions? These questions are explored in the following sections.

3.2 Embodiment

The impact of embodiment and social behaviour for children learning English as their L1 has been explored in a laboratory setting. Neither [24] nor [71] found significant differences due to the embodiment of the robot in their studies on children's vocabulary acquisition. However, this may be due in part to methodological limitations. Gordon et al. [24] only found an average of one word learned per interaction, leaving very little room for observing differences; similarly [71] only compared the learning of six words. These studies were conducted with children between the ages of 3 and 8 years. The relatively small gains are therefore quite surprising, due to the speed at which children at this age acquire language [40]. Given the non-significant results or the small effect sizes in these studies, it is difficult to draw conclusions on what could make robot language tutoring effective.

Rosenthal-von der Pütten et al. [58] found that language alignment, i.e., the use of similar verbal patterns between interacting parties, when using an L2 appears to not be affected when using a virtual robot as opposed to a real one. Participants completed a pre-test and were then invited for a second session at a later date. During the second session the participants were asked to play a guessing game with an

agent, either the real NAO robot or a virtual representation of one. The study reported whether the participants used the same words as the agent, but no significant difference was found. This may be due to some issues with the experimental design: the authors suggest the post-test was given straight after a relatively long session with the agent, and participants may have been fatigued.

Moriguchi et al. [49] looked at age differences for young children and how they learned from a robot compared to a person. Children between the ages of 4 and 5 years were taught using an instructional video: one group of children was shown a video in which a human taught them new words, while another group of children was shown a video with the same material, but using a robot tutor. While children aged 5 were able to perform almost as well when taught by a robot, those aged 4 did not seem to learn from the robot at all. It is unknown as to whether this result would transfer to the use of a physically-present robot, rather than one shown on a video screen.

These studies above do not provide support that the mere physical presence of the robot has an advantage for language learning. However, there is evidence for the physical presence of a robot having a positive impact on various interaction outcomes, including learning [46]. The lack of a clear effect of a physical robot on language learning might be due to a scarcity of experimental data. However, it is also likely that the effectiveness of robot tutors lies not in their physical presence, but instead in the social behaviour that a robot can exhibit and the motivational benefits this carries. This is explored in the next section.

3.3 Social Behaviour

Social behaviour has previously been studied in the context of children learning languages. Saerbeck et al. [60] explored the impact of 'socially supportive' behaviours on child learning of the Toki Pona language, using an iCat robot as a tutor. These behaviours included verbal and non-verbal manipulations which aimed to influence feedback provision, attention guiding, empathy, and communicativeness. It was found that the tutor with these socially supportive behaviours significantly increased the child's learning potential when compared to a neutral tutor. This study used a variety of measures including vocabulary acquisition, as other studies have, but also included pronunciation and grammar tests. Another study which did not only consider vocabulary acquisition was [26]. French and Latin verb conjugations were taught by a NAO robot to children aged 10 to 12 years old. In one condition, the robot would look towards the student whilst they completed worksheets, but in the other, the robot would look away. Although gaze towards the child was predicted to lead to greater social facilitation effects, and therefore higher performance, this was not observed.

Kennedy et al. [36] investigated the effects of verbal immediacy on the effect of learning in children. A NAO was used to teach French to English-speaking children in a task involving the gender of nouns and the use of articles 'le' and 'la'. A high verbal immediacy condition was designed in which the robot would exhibit several verbal immediacy behaviours, for example calling the child by name, providing positive feedback, and asking children how they felt about their learning. When contrasted with a robot without this behaviour, no significant learning differences were observed. However, children showed significant improvement in both conditions when comparing pre- and post-test scores, and were able to retain their acquired knowledge as measured by means of a retention test. This suggests that the particularities of robot behaviour do not manifest themselves in the short-term, but could be potentially be observed over the longer term.

In [2], a robot acted as a teaching assistant for the purpose of teaching English to Iranian students. A survey found that students who were taught by the robot were significantly less anxious about their lessons than those that were not. This was thought to be due to a number of factors, including the fact that the robot was programmed to make intentional mistakes which the students could correct, which could have made students less concerned about their own mistakes.

3.4 Summary

In summary, promising results have been found for the use of robots as constrained language tutors for children and adults, with the presence of the robot improving learning outcomes [1,2,33,64]. However, the impact of robot embodiment in this context has not been explored in depth, leaving an important question largely unanswered: do robots hold an advantage over tablets or virtual characters for language tutoring? The impact of social behaviour is also less clear, with some positive results [60], but also inconclusive results [26]. Robots open up new possibilities in teaching that were previously unavailable, such as the robot taking the role of a peer. By having an agent that is less like a teacher and more like a peer, anxiety when learning a new language could be reduced [2]. Despite an increasing interest, there are still relatively few studies that have considered robot language tutoring, leaving space to explore novel aspects of language learning.

4 Designing Robot Tutoring Interactions for Children

Several design issues with respect to robot-guided L2 tutoring have to be considered before an evaluation of robot-child tutoring success is possible. In particular, multiple design choices have to be considered to create pleasant, challeng-

ing, and pedagogically sound interactions between robot and child [69]. First, we will discuss pedagogical issues that ensure optimal conditions for language learning. Second, we will present various design issues specifically relating to the child–robot interactions. Finally, we will discuss how to manage personalised interactions during tutoring. The section builds on some related work as well as various studies conducted in the context of the L2TOR project.

4.1 Pedagogical Issues

It is imperative to understand how previous research findings can be put into practice to support successful L2 acquisition. Although the process of language learning does not drastically differ between L1 and L2, there are a few notable differences as we already discussed in Sect. 2. For the L2TOR project a series of pedagogical guidelines was formulated, based on existing literature and pilot data collected within our project. These guidelines concern: (a) age differences; (b) target word selection; (c) the use of a meaningful context and interactions to actively involve the child; and (d) the dosage of the intervention. These specific aspects were chosen based on a review of the literature showing that they are the most crucial factors to consider in designing an intervention for language teaching in general and specifically L2 (see e.g., [29,51]).

4.1.1 Age Effects

From what age onward can we use social robots to support L2 learning effectively? From a pedagogical point of view, it is desirable to start L2 tutoring as early as possible, especially for children whose school language is an L2, because this could bridge the gap in language proficiency that they often have when entering primary school [29]. Various studies have targeted children as young as 3 years focusing on interactive storytelling in the L1 [22] or on L2 tutoring [73]. However, preschool-aged children (3 to 5 years old) undergo major cognitive, emotional and social developments, such as the expansion of their social competence [15]. So, whereas older children may have little difficulty engaging in an interaction with a robot, younger children may be more reliant on their caregivers or show less engagement in the interaction. Therefore, we may expect that child–robot interactions at those ages will also present some age-related variation. Clarifying these potential age differences is essential as, in order to be efficient, interactive scenarios with robots must be tailored to the diverging needs of children.

In [6], we sought to determine whether there are age-related differences in first-time interactions with a peer-tutor robot of children who have just turned 3 and children who are almost 4 years old. To this end, we analysed the engagement of 17 younger children (M_{age} = 3.1 years, SD_{age} = 2 months) and 15 older children (M_{age} = 3.8 years, SD_{age} = 1 month) with a NAO robot as part of the larger feedback experiment discussed in Sect. 4.2.6. These children first took part in a group introduction to familiarise them with the NAO robot; a week later they had a one-on-one tutoring session with the robot. We analysed the introductory part of this one-on-one session, which consisted of greeting, bonding with, and counting blocks with the robot. All speech was delivered in Dutch, except for the target words (i.e., 'one', 'two', 'three', and 'four'), which were provided in English. We analysed the children's engagement with the robot as measured through eye-gaze towards the task environment (robot and blocks) compared to their gazes outside the task environment (experimenter, self, and elsewhere), as this is suggested to indicate how well the child is "connected" with the task [62].

In short, the analyses revealed that the older children gazed significantly longer towards the robot than the younger children, and that the younger children spent more time looking elsewhere than the older children. Moreover, the average time the older children maintained each gaze towards the robot was longer than that of the younger children.

It is possible that the 3-year-olds have trouble being engaged with a language learning task, but it may also be that the NAO robot is somewhat intimidating for 3-year-olds. As such, for them either group interactions [22] or a more "huggable" robot (e.g., Tega) [73] could be more appropriate. Moreover, [49] also found children at the age of 5 years to be more responsive to robot tutoring. Drawing from these findings about 3-year-olds, combined with experiences from other pilots with 4- and 5-year-olds, we decided to develop the L2TOR tutoring system for 5-year-olds, as they generally appear to feel more comfortable engaging one-on-one with the robot than 3- and 4-year-olds.

4.1.2 Target Words

Another important aspect to consider is what words are taught. Previous research recommends that vocabulary items should be taught in semantic clusters and embedded in a conceptual domain [11,51]. For L2TOR, three domains were chosen: (a) number domain: language about basic number and pre-mathematical concepts; (b) space domain: language about basic spatial relations; and (c) mental states domain: language about mental representations such as 'being happy' and propositional attitudes such as 'believe' or 'like'. These domains were selected for their feasibility, as well as their relevance and applicability in L2 tutoring sessions in a preschool setting. Appropriate words to be taught for each domain are words that children should be familiar with in their L1, as the goal of the intervention is not to teach children new mathematical, spatial, and mental state concepts, but rather L2 labels for familiar concepts in these three domains. This will

enable children to use their L1 conceptual knowledge to support the learning of L2 words. To select appropriate target words and expressions that children are familiar with in their L1, a number of frequently used curricula, standard tests, and language corpora were used. These sources were used both for identifying potential targets, and for checking them against age norms to see whether they were suitable for the current age group (for more details, see [53]). Thus, target words selection should be based both on semantic coherence and relevance to the content domain and on children's L1 vocabulary knowledge.

4.1.3 Meaningful Interaction

An additional aspect of L2 teaching is the way in which new words are introduced, which may come to affect both learning gains as well as the level of engagement. Research has indicated that explicit instruction on target words in meaningful dialogues involving defining and embedding words in a meaningful context yields higher word learning rates than implicit instruction through fast mapping (i.e., mapping of a word label on its referent after only one exposure) or extracting meaning from multiple uses of a word in context as the basic word learning mechanisms [48,51]. Therefore, for the L2TOR project, an overall theme for the lessons was selected that would be familiar and appealing to most children, and, as such, increase childrens engagement during the tutoring sessions. This overall theme is a virtual town that the child and the robot explore together, and that contains various shops, buildings, and areas, which will be discovered one-by-one as the lesson series progresses. All locations are familiar to young children, such as a zoo and a bakery. During the lessons, the robot and the child discover the locations, and learn L2 words by playing games and performing simple tasks (e.g., counting objects or matching a picture and a specific target word). The child and the robot are awarded a star after each completed session, to keep children engaged in the tasks and in interacting with the robot. Thus, the design chosen for L2TOR is thought to facilitate higher learning gains as it involves explicit teaching of target words in a dialogue taking place in a meaningful context. Moreover, this design should facilitate engagement as it involves settings that are known and liked by children.

4.1.4 Dosage of Language Input

The final pedagogical aspect that was identified in the literature concerns the length and intensity, or dosage, of the intervention. Previous research has shown that vocabulary interventions covering a period of 10 to 15 weeks with one to four short 15- to 20-min sessions per week are most effective. As for the number of novel words presented per session, the common practice is to offer 5 to 10 words per session, at least in L1 vocabulary interventions [47]. However, not much is known about possible differences between L1 and L2 interventions with regard to this aspect. Therefore, to determine the number of target words to be presented in the L2TOR project lesson series, a pilot study was conducted. In this study, we taught English words to one hundred 4- and 5-year-old Dutch children with no prior knowledge of English. We started by teaching the children 10 words; when these were established, more words were added. The results showed that, for children to learn any of these words at all, the maximum number of L2 words that could be presented in one session was six. We also found that a high number of repeated presentations of each word was necessary for word learning: each word in our study was presented 10 times. Yet, children's accuracy rates in the translation and comprehension tasks in our study were lower than in earlier work on L1 learning. A possible explanation might be that the items included in the study were relatively complex L2 words (e.g., adjectives like 'empty') rather than concrete nouns such as 'dog' or 'house'. These items are probably more difficult for children who had no prior exposure to the target language. However, within the L2TOR project the choice was made to include these relatively complex items given their relevance for L2 learning within an academic context [52]. Thus, it was decided that in all the lessons included within the L2TOR project a maximum of six words will be presented in each lesson and each word will be repeated at least ten times throughout the lesson.

4.2 Child–Robot Interaction Issues

Not only pedagogical issues need to be considered when designing a social robot tutor, but also other issues relating to how the interactions between the robot and child should be designed. As mentioned, we focus on how to design the interactions to be pleasant, challenging, and pedagogically sound. In this section, we discuss six aspects that we deem important: (a) first encounters; (b) the role of the robot; (c) the context in which the interactions take place; (d) the non-verbal behaviours and (e) verbal behaviours of the robot; and (f) the feedback provided by the robot.

Before elaborating on these guidelines, it is important to remind the reader that in L2TOR, we are designing the robot to operate fully autonomously. Ideally, this would include the possibility to address the robot in spoken language and that the robot can respond appropriately to this. However, as previously mentioned, current state-of-the-art in speech recognition for child speech does not work reliably. Kennedy et al. [37] compared several contemporary ASR technologies and have found that none of them achieve a recognition accuracy that would allow for a reliable interaction between children and robots. We have therefore decided to mediate the interactions using a tablet that can both display the learning context (e.g., target objects) and monitor children's responses

to questions. This has the consequence that the robot cannot monitor children's L2 production autonomously, but it can monitor children's L2 comprehension through their performance with respect to the lesson content presented on the tablet.

4.2.1 Introducing the Robot

The first encounter between robot and child plays a large role in building the child's trust and rapport with the robot, and to create a safe environment [72], which are necessary to facilitate long-term interactions effectively. For example, [21] has shown that a group introduction in the kindergarten prior to one-on-one interactions with the robot influenced the subsequent interactions positively. Moreover, [72] have shown that introducing the robot in a one-to-many setting was more appreciated than in a one-on-one setting, because the familiarity with their peers can reduce possible anxiety in children.

We, therefore, developed a short session in which the robot is introduced to children in small interactive groups. In this session, the experimenter (or teacher) first tells a short story about the robot using a picture book, explaining certain similarities and dissimilarities between the robot and humans in order to establish some initial common ground [14,33]. During this story, the robot is brought into the room while in an animated mode (i.e., turned on and actively looking around) to familiarise the children with the robot's physical behaviour. The children and the robot then jointly engage in a meet-and-greet session, shaking hands and dancing together. We observed in various trials that almost all children were happy to engage with the robot during the group session, including those who were a bit anxious at first, meaning these children likely benefited from their peers' confidence. Although we did not test this experimentally, our introduction seems to have a beneficial effect on children's one-on-one interaction with the robot.

4.2.2 Framing the Robot

One of the questions that arises when designing a robot tutor is: How should the robot be framed to children, such that interactions are perceived to be fun, while at the same time be effective to achieve language learning? We believe it is beneficial to frame the robot as a peer [5,7,24], because children are attracted to various attributes of a robot [33] and tend to treat a robot as a peer in long-term interactions [64]. Moreover, framing the robot as a peer could make it more acceptable when the flow of the interaction is suboptimal due to technical limitations of the robot (e.g., the robot being slow to respond or having difficulty interpreting children's behaviours). In addition, framing the robot as a peer who learns the new language together with the child sets the stage for learning by teaching [64].

While the robot is framed as a peer and behaves like a friend of the child, the tutoring interactions will be designed based on adult-like strategies to provide the high quality input children need to acquire an L2 [39], such as providing timely and sensible non-verbal cues or feedback. So, in L2TOR we frame the robot as a peer, it behaves like a peer, but it scaffolds the learning using adult-like teaching strategies.

4.2.3 Interaction Context

To facilitate language learning, it is important to create a contextual setting that provides references to the target words to be learned. The embodied cognition approach, on which we base our project, states that language is grounded in real-life sensorimotor interactions [28], and consequently predicts that childrens interactions with real-life objects will benefit vocabulary learning [23]. From this approach, one would expect children to learn new words better if they manipulate physical objects rather than virtual objects on a tablet, as the former allows children to experience sensorimotor interactions with the objects. However, for technical reasons discussed earlier, it would be convenient to use a tablet computer to display the context and allow children to interact with the objects displayed there. The question is whether this would negatively affect learning. Here, we summarise the results from an experiment comparing the effect of real objects versus virtual objects on a tablet screen on L2 word learning [68]. The main research question is whether there is a difference in L2 vocabulary learning gain between children who manipulate physical objects and children who manipulate 3D models of the same objects on a tablet screen.

In this experiment, 46 Dutch preschoolers ($M_{age} = 5.1$ years, $SD_{age} = 6.8$ months; 26 girls) were presented with a story in Dutch containing six L2 (English) target words (i.e., 'heavy', 'light', 'full', 'empty', 'in front of,' and 'behind'). These targets were chosen as children should benefit from sensorimotor interactions with objects when learning them. For example, learning the word 'heavy' could be easier when actually holding a heavy object rather than seeing a 3D model of this object on a tablet screen. Using a between-subjects design, children were randomly assigned to either the tablet or physical objects condition. During training, the target words were each presented ten times by a human. Various tests were administered to measure the children's knowledge of the target words, both immediately after the training and one week later to measure children's retention of the target words.

Independent-samples t-tests revealed no significant differences between using a tablet or physical objects on any of the tasks, as indicated by childrens mean accuracy scores on the direct and delayed post-tests (see Fig. 1; all p val-

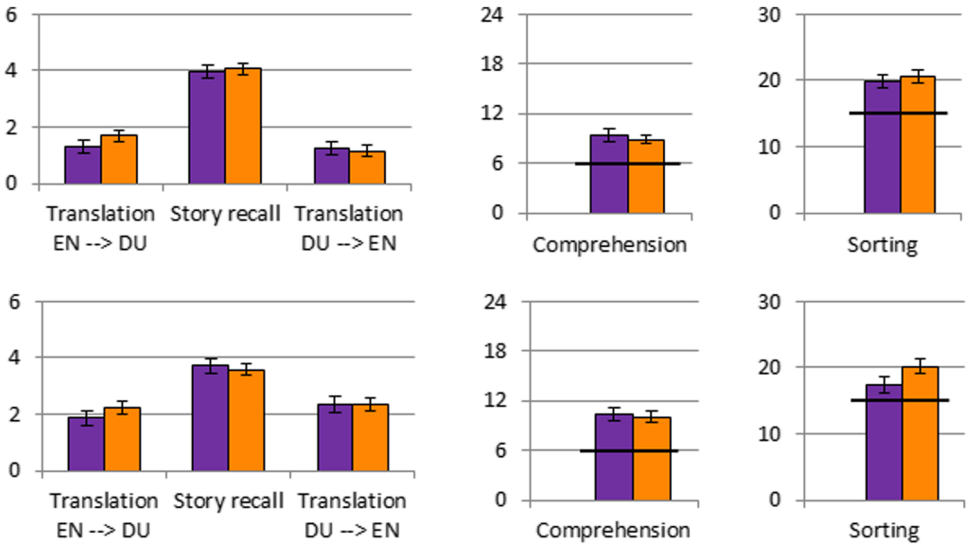


Fig. 1 Mean accuracy scores on the direct post-test (top) and the delayed post-test (bottom). Purple bars refer to the object condition; orange bars to the tablet condition. Reprinted from [68]. (Color figure online)

ues $>$ .243). In the receptive tests (the comprehension task and sorting task), children scored significantly above chance level (indicated by the black line), irrespective of condition (all p values $<$ .001). Interestingly, in both conditions, the mean scores on the Dutch-to-English translation task were higher for the delayed post-test than for the immediate post-test (both p values $<$ .001), possibly indicating some sort of "sleep effect". These findings indicate that it does not matter much whether the context is presented through physical objects or a tablet computer.

Displaying the context (i.e., target objects) on a tablet does not seem to hamper learning, which is convenient, since using a tablet makes designing contexts more flexible and reduces the need to rely on complex object recognition and tracking. Because of this, the lessons in the L2TOR project are displayed on a tablet, which is placed between the child and the robot (see Fig. 2). This tablet not only displays the target objects (e.g., a set of elephants in a zoo), but also allows children to perform actions on these objects (e.g., placing a given number of elephants in their cage). Since at present ASR for children is not performing reliably [37], the robot cannot monitor children's pronunciation or other verbal responses. We therefore focus on language comprehension rather than language production and use the tablet to monitor comprehension. The use of a tablet in the interaction allows us to monitor the child's understanding of language and to control the interaction between child and robot.

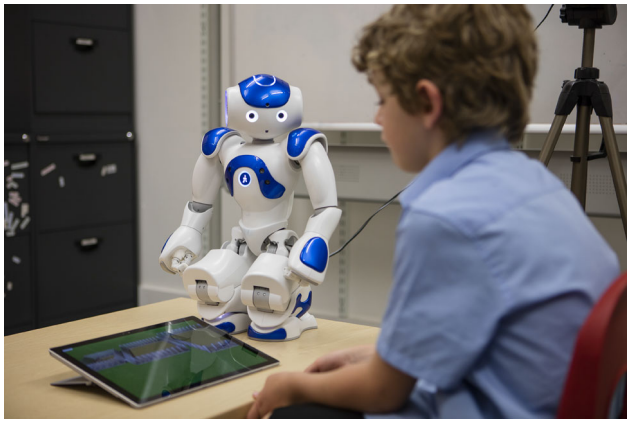


Fig. 2 The L2TOR setup includes the NAO robot standing to the side of the child with a tablet in between them

4.2.4 Non-verbal Behaviour

Human language production is typically accompanied by non-verbal cues, such as gestures or facial expressions. It is therefore not surprising that research in children's language development has shown that the use of gestures facilitates L2 learning in various ways (e.g., [25,59,65]). Gestures could take the form of deictic gestures, such as pointing to refer to physical objects near the child, or of iconic gestures used to emphasize physical features of objects or actions in a more representational manner. Such iconic gestures help to build congruent links between target words and perceptual or motor information, so learners may benefit not only from observing gestures, but also by way of execution, such as enactment and imitation [23,25].

Due to its physical presence in the child's referential world, a robot tutor has the ability to use its physical embodiment to its advantage when interacting with the child, for example, through the manipulation of objects in the real world, or simply through the use of gestures for various communicative purposes. We believe that the robot's ability to use gestures is one of the primary advantages of a robot as tutor compared to a tablet computer, since it can enrich the language learning environment of the child considerably by

exploiting the embodiment and situatedness of the robot to facilitate the child's grounding of the second language.

Even though a growing body of evidence suggests that non-verbal cues, such as gestures aid learning, translating human's non-verbal behaviour to a robot like NAO remains a challenge, mostly due to hardware constraints. For instance, the NAO robot is limited by its degrees of freedom and constraints with respect to its physical reach, making it unable to perform certain gestures. Motions may sometimes seem rigid, causing the robot's movements to appear artificial rather than human-like. Especially when certain subtleties are required when performing a gesture, such shortcomings are not desirable. A noteworthy complication comes with the NAO's hand, which has only three fingers that cannot move independently of one another. This makes an act such as finger-counting, which is often used for the purpose of explaining numbers or quantities, practically impossible.

This, thus, requires a careful design and testing of appropriate referential gestures, because otherwise they may harm learning [35].

4.2.5 Verbal Behaviour

One potential advantage of using digital technologies, such as robots, is that they can be programmed to speak multiple languages without an accent. However, NAO's text-to-speech engines do generate synthetic voices and have few prosodic capacities. Yet, studies have shown that children rely on prosodic cues to comprehend spoken language (e.g., [16]). Moreover, adults typically use prosodic cues to highlight important parts of their speech when addressing children. In addition, the lack of facial cues of the NAO robot may potentially hinder the auditory-visual perception processes of both hearing-impaired and normal-hearing children [19]. These limitations pose the question to what extent children can learn the pronunciation of L2 words sufficiently well.

To explore this, a Wizard-of-Oz (WoZ) experimental pilot was devised using the NAO robot and a tablet for tutoring and evaluating English children counting up to five in German. The task involved multiple steps to gradually teach children to count, in L2, animals shown on screen. First, the robot-tablet concept was introduced, with the robot describing content displayed on the tablet screen, and the children were trained on how and when to provide answers by means of touching images on said screen. The children then proceeded with the main task, which involved the counting of animals, first in English and later in German. The interaction was managed by using multiple utterances from a WoZ control panel in order to prompt the children to give the answer only after they were asked to. The WoZ operator triggered appropriate help and feedback from the robot to the child when required. Finally, at the end of the task, the robot asked the children to count up to five again with the robots help and then without any help at all. The purpose of this step was to evaluate whether the children were able to remember the pronunciation of the German numbers and if they were able to recall them with no support.

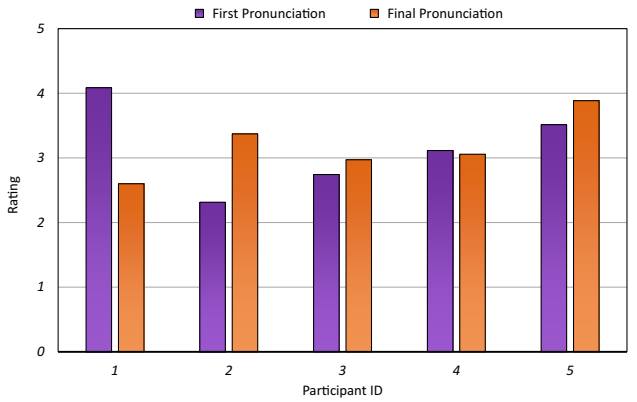


Fig. 3 Pronunciation ratings from seven German native speakers for 5 child participants. Three of the children improve over the course of the interaction, although one child has initially accurate pronunciation that drops over time, possibly due to fatigue

Voice and video recordings were used to record the interactions with five children aged 4 to 5 years old. The first and final repetitions of the children pronouncing the German words were recorded and rated for accuracy on a 5-point Likert scale by seven German-native coders; intraclass correlation $ICC(2,7) = .914$, indicating "excellent" agreement [13]. Based on these ratings, our preliminary findings are that repetitions generally improve pronunciation. Several children initially find it hard to pronounce German numbers but they perform better by the end (Fig. 3). This may be because some children had trouble recalling the German numbers without help. We believe that the task needs updating to improve the children's recall (by, for example, including more repetitions). In addition, it should be noted that children generally find it difficult to switch from English to German.

To conclude, children can learn the pronunciation of the L2 from the robot's synthetic voice, but we should compare this to performance ratings of children that have learned the L2 from native speakers. It is worth noting that they seem to have some reservation speaking a foreign language, but whether or not this is due to the presence of the robot is unknown.

4.2.6 Feedback

A typical adult-like strategy known to support language learning is the use of appropriate feedback [3]. Adult caregivers tend to provide positive feedback explicitly (e.g., 'well done!') and negative feedback implicitly by recasting the correct information (e.g., 'that is the rabbit, now try again to touch the monkey'). However, evidence suggests that a peer

does not generally provide positive feedback and that they provide negative feedback explicitly without any correction (e.g., 'no, that is wrong!'). So, when the robot is framed as a peer, should it also provide feedback like a peer?

To explore this, we carried out an experiment to investigate the effect the type of feedback has on children's engagement [17,18]. In the experiment, sixty-five 3-year-old children (30 boys, 35 girls; M_{age} = 3.6 years, SD_{age} = 3.6 months) from different preschools in the Netherlands participated. Six children stopped with the experiment before it was finished and were excluded from the data. The children were randomly assigned to one of three conditions, varying the type of feedback: adult-like feedback, peer-like feedback, and no feedback. The adult-like feedback of the robot used reformulations to correct the children in case they made a mistake (e.g., 'three means *three*', where the text in italics represents what the robot said in the L2, here English; the rest was said in the L1, here Dutch) and positive feedback ('well done!') when children responded correctly. In the peer-like condition, only explicit negative feedback without correction was provided whenever children made a mistake ('that is wrong!') and no feedback was provided when they responded correctly. In the no feedback condition, the robot simply continued with the next task without providing any feedback.

During the experiment, the robot taught the native Dutch-speaking children counting words one to four in English. The interaction consisted of an introductory phase followed by the tutoring phase. During the introductory phase, the target words (i.e., 'one', 'two', 'three', and 'four') were described and associated with their concept in sentences such as 'I have *one* head', 'I have *two* hands', 'I have *three* fingers', and 'there are *four* blocks'. We analysed the introductory phase as part of the age-effects study reported in Sect. 4.1.1. In the tutoring phase, the robot asked the child to pick up a certain number of blocks that had been placed in front of them. All instructions were provided in Dutch and only the target words were provided in English. After the child collected the blocks, the robot provided either adult-like feedback, peer-like feedback, or no feedback depending on the experimental condition assigned to the child.

As a result of the relatively low number of repetitions of the target words over the course of the interaction, we did not expect to find any effects with respect to learning gain. However, the objective was not to investigate the effect feedback has on learning, but rather on the child's engagement with the robot as an indicator of learning potential [12]. As for the age-effect study, we analysed engagement by annotating the children's eye-gaze towards the robot, human experimenter, to the blocks, and elsewhere, and measured the average time children maintained their gaze each time they looked at one of these targets.

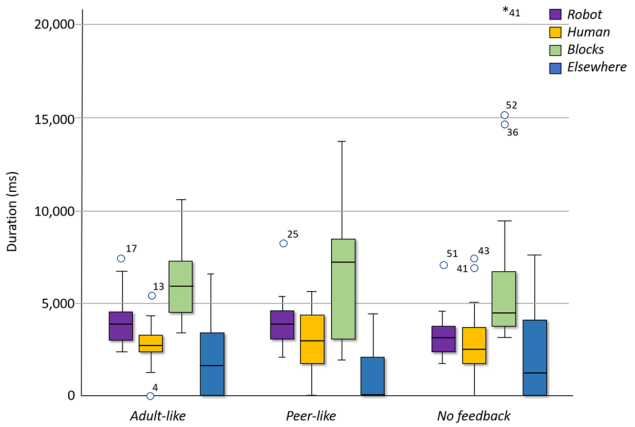


Fig. 4 Mean duration per gaze to the robot, blocks, experimenter, and elsewhere for the three feedback conditions

Results from a repeated measures ANOVA indicated that, on average, the children maintained their gaze significantly longer at the blocks and the robot than at the experimenter, regardless of their assigned condition (see Fig. 4).

However, we did not see any significant differences in the gaze duration across the three conditions. As such, the way the robot provides feedback does not seem to affect the engagement of the child with the robot. This would suggest that, as far as the child's engagement with the robot and task is concerned, it does not matter how the robot provides feedback or whether the robot provides feedback at all. Hence, the choice for the type of feedback that the robot should give can, thus, solely be based on the effect feedback has on learning gain. Future work will investigate which type of feedback is most effective for learning.

4.3 Interaction Management

4.3.1 Objective

To realise robot-child tutoring interactions that provide a pleasant and challenging environment for the child, while at the same time being effective for L2 learning, interaction management plays a crucial role.

As children typically lose interest when a lesson is either too easy or too difficult, personalisation of the lessons to each child's performance is very important. The tutor has to structure the interaction, needs to choose the skills to be trained, must adjust the difficulty of the learning tasks appropriately, and has to adapt its verbal and non-verbal behaviour to the situation. The importance of personalised adjustments in the robot's behaviour has been evidenced in research showing that participants who received personalised lessons from a robot outperformed others who received non-personalised training [5,45]. Suboptimal robot behaviour (e.g., too much, too distracting, mismatching, or in other ways inappropriate)

can even hamper learning [35]. Therefore lessons should be adapted to the knowledge state (i.e., level) of the child [70].

Along these lines, the L2TOR approach is to personalise language tutoring in HRI by integrating knowledge-tracing into interaction management [61]. This adaptive tutoring approach is realised in a model of how tutors form mental states of the learners by keeping track of their knowledge state and selecting the next tutoring actions based on their likely effects on the learner. For that purpose, an extended model based on Bayesian Knowledge Tracing was built that combines knowledge tracing (what the learner learned) and tutoring actions in one probabilistic model. This allows for the selection of skills and actions based on notions of optimality: the desired learner's knowledge state as well as optimal task difficulty.

4.3.2 Proposed Model

A heuristic is employed that maximises the beliefs of all skills while balancing the single skill-beliefs with one another. This strategy is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system [43]. For the choice of actions, the model enables simulation of the impact each action has on a particular skill. To keep the model simple, the action space only consists of three different task difficulties (i.e., easy, medium, hard).

4.3.3 Results

As an evaluation, the model was implemented and tested with a robot language tutor during a game-like vocabulary tutoring interaction with adults ($N = 40$) [61].

We adopted the game 'I spy with my little eye'. In this game, the NAO robot describes an object which is displayed on a tablet along with some distractors, by referring to its descriptive features in an artificial L2 (i.e., "Vimmi"). The student then has to guess which object the robot refers to. The overall interaction structure, consisting of five phases (i.e., opening, game setup, test-run, game, closing), as well as the robot's feedback strategies were based on our observations of language learning in kindergartens. After the tutoring interaction, a post-test of the learned words was conducted.

The results revealed that learners' performance improved significantly during training with the personalised robot tutor (Fig. 5). A mixed-design ANOVA with training phase as a within-subjects factor and training type as between-subject factor demonstrated a significant main effect of training phase ($F(1, 38) = 66.85$, $p < .001$, $\eta^2 = .64$), such that learners' performance was significantly better in the final phase as compared to the initial phase. Crucially, participants who learned in the adaptive condition had a higher number of correct answers as compared to the control condition ($F(1, 38) = 6.52$, $p = .02$, $\eta^2 = .15$). Finally, the

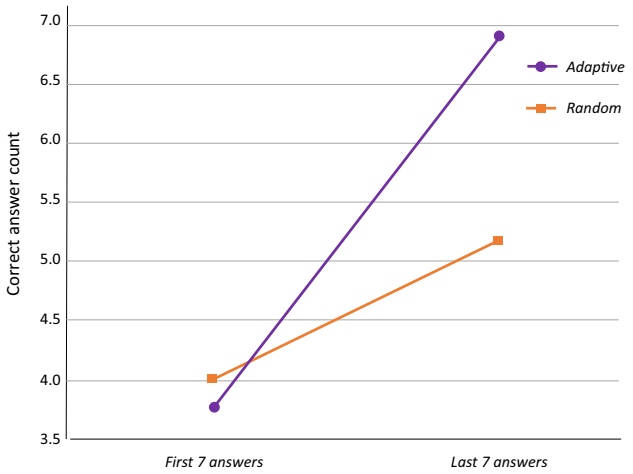


Fig. 5 Mean numbers of correct answers at the beginning (first 7) and end (last 7) of the interaction in the different conditions. Adapted from [61]

Table 1 Results of both post-tests (L1-to-L2 and L2-to-L1): Means (M) and standard deviation (SD) of correct answers grouped by the experimental conditions

	Adaptive (A)		Control (C)	
	M	SD	M	SD
L1-to-L2	3.95	2.56	3.35	1.98
L2-to-L1	7.05	2.56	6.85	2.48

Adapted from [61]

interaction between training phase and type was also significant ($F(1, 38) = 14.46$, $p = .001$, $\eta^2 = .28$), indicating that the benefit of the adaptive training developed over time.

The results of the post-test did not show significant differences between the two conditions, which may be due to the way in which responses were prompted during the training sessions and post-test (Table 1). In the training sessions participants saw pictures relating to the meaning of the to-be-learned words, whereas in the post-test they received a linguistic cue in form of a word they had to translate. Although no main effect of training type emerged in the post-test, some details are nevertheless worth mentioning. In the L1-to-L2 post-test, a maximum of ten correct responses was achieved by participants of the adaptive-model condition, whereas the maximum in the control condition were six correct answers. Moreover, there were two participants in the control condition who did not manage to perform any L1-to-L2 translation correctly, while in the adaptive-model condition, all participants achieved at least one correct response (see Fig. 6).

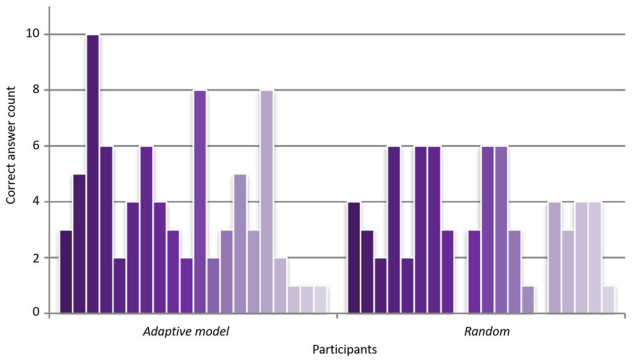


Fig. 6 Participant-wise amount of correct answers grouped by the different conditions for the L1-to-L2 post-test. Adapted from [61]

4.3.4 Outlook

This basic adaptive model will be extended by further integrating skill interdependencies as well as affective user states. Both have already been shown to improve learning [34,63]. In addition, the model can, and is meant to, provide a basis for exploiting the full potential of an embodied tutoring agent, and will therefore be advanced to the extent that the robot's verbal and non-verbal behaviour will adapt to the learner's state of knowledge and progress. Specifically, it aims to enable dynamic adaptation of (a) embodied behaviour such as iconic gesture use, which is known to support vocabulary acquisition as a function of individual differences across children (cf. [59]); (b) the robot's synthetic voice to enhance comprehensibility and prosodic focusing of content when needed; and (c) the robot's socio-emotional behaviour depending on the learners' current level of motivation and engagement.

5 Evaluation Framework for Robot L2 Tutoring

In this section, we discuss our plans for evaluating our robot-assisted L2 vocabulary intervention. While this section describes future plans rather than already completed work, it also provides guidelines for evaluating tutoring systems similar to the L2TOR system. The first step in an evaluation is the development of pre- and post-tests designed to assess children's learning of the targeted vocabulary through comprehension and translation tasks, as well as tasks assessing deep vocabulary knowledge (i.e., conceptual knowledge associated with a word). Not directly targeted but semantically-related vocabulary will also be assessed, as well as general vocabulary and other skills related to word learning (e.g., phonological memory). This is important as children learn not only the words directly used, but can also use these words to bootstrap their further vocabulary learning in the same as well as related domains [51].

In addition to assessing children's L2 word learning, we will evaluate the word learning process during the interactive sessions between children and the robot by observing, transcribing, and coding video-taped interactions. Measures will include children's and the robot's participation and turn-taking, the type of questions, recasts and expansions, the semantic contingency of responses and expansions, and the coherence and length of episodes within the sessions. All these aspects are known to promote language learning [9,44]. Therefore, it is important to evaluate how these processes are taking place within the context of language learning with a social robot.

Finally, given the importance of motivation, we will observe how children comply with the robot's initiatives and instructions, how involved they are in the intervention, and to what extent they express positive emotions and well-being during the lessons [41]. The intervention will consist of multiple sessions, such that children's learning, motivation, and interaction with a social robot can be judged over time.

The design of the evaluation study will be based on a comparison between an experimental and a control group. The experimental group will be taught using the social robot while the control group will receive a placebo training (e.g., non-language activity with the robot). This design is very common in educational research as it enables testing whether children who participate in an educational programme (L2TOR in this case) learn more or just as much as children who follow the normal curriculum. Additionally, learning gains with the robot will be compared to learning gains using an intelligent tutoring system on a tablet, to test the additional value of a social robot above existing technology used in education. In evaluating the robot-supported program developed within L2TOR, our aim is not only to assess the effectiveness of the specific tutoring by the L2TOR robot, but also to provide recommendations for further technological development and guidelines for future use of social robots in (L2) language tutoring situations.

6 Conclusion

In this paper, we have presented guidelines for designing social robots as L2 tutors for preschool children. The guidelines cover a range of issues concerning the pedagogy of L2 learning, child–robot interaction strategies, and the adaptive personalisation of tutoring. Additional guidelines for evaluating the effectiveness of L2 tutoring using robots were presented.

While the benefits of social robots in tutoring are clear, there are still a range of open issues on how robot tutors can be effectively deployed in educational settings. The spe-

cific focus of this research programme –tutoring L2 skills to young children– requires an understanding of how L2 learning happens in young children and how children can benefit from tutoring. Transferring the tutoring to social robots has highlighted many questions: should the robot simulate what human tutors do? Should the robot be a peer or a teacher? How should the robot blend L1 and L2? How should feedback be given?

Our aim is to develop an autonomous robot: this incurs several complex technical challenges, which cannot currently be met by state-of-the-art AI and social signal processing. ASR of child speech, for example, is currently insufficiently robust to allow spoken dialogue between the robot and the young learner. We propose a number of solutions, including the use of a tablet as an interaction-mediating device.

Our and our colleagues' studies show that social robots hold significant promise as tutoring aids, but a complex picture emerges as children do not just learn by being exposed to a tutoring robot. Instead, introducing robots in language learning will require judicious design decisions on what the role of the robot is, how the child's learning is scaffolded, and how the robot's interaction can support this.

Acknowledgements We are grateful to all research assistants, participating schools, parents, and their children for their assistance in this project.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Funding The L2TOR project is funded by the H2020 Framework Programme of the EC, Grant Number: 688014.

References

1. Alemi M, Meghdari A, Ghazisaedy M (2014) Employing humanoid robots for teaching english language in Iranian Junior High-Schools. Int J Hum Robot 11(3). https://doi.org/10.1142/S0219843614500224
2. Alemi M, Meghdari A, Ghazisaedy M (2015) The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. Int J Social Robot. pp 1–13. https://doi.org/10.1007/s12369-015-0286-y
3. Ateş-Şen AB, Küntay AC (2015) Childrens sensitivity to caregiver cues and the role of adult feedback in the development of referential communication. Acquis Ref 15:241–262
4. Barcelona European Council (2002) Presidency Conclusions, part I, 43.1
5. Baxter P, Ashurst E, Read R, Kennedy J, Belpaeme T (2017a) Robot education peers in a situated primary school study: personalisation promotes child learning. PLoS One 12(5):e0178126
6. Baxter P, de Jong C, Aarts A, de Haas M, Vogt P (2017b) The effect of age on engagement in preschoolers child–robot interactions. In: Companion proceedings of the 2017 ACM/IEEE international conference on HRI
7. Belpaeme T, Baxter PE, Read R, Wood R, Cuayáhuitl H, Kiefer B, Racioppa S, Kruijff-Korbayová I, Athanasopoulos G, Enescu V et al (2012) Multimodal child–robot interaction: building social bonds. J Hum–Robot Interact 1(2):33–53
8. Belpaeme T, Kennedy J, Baxter P, Vogt P, Krahmer EJ, Kopp S, Bergmann K, Leseman P, Küntay AC, Göksun T, Pandey AK, Gelin R, Koudelkova P, Deblieck T (2015) L2TOR—second language tutoring using social robots. In: Proceedings of first international workshop on educational robots
9. Bornstein MH, Tamis-LeMonda CS, Hahn CS, Haynes OM (2008) Maternal responsiveness to young children at three ages: longitudinal analysis of a multidimensional, modular, and specific parenting construct. Dev Psychol 44(3):867
10. Brojde CL, Ahmed S, Colunga E (2012) Bilingual and monolingual children attend to different cues when learning new words. Front Psychol 3:155. https://doi.org/10.3389/fpsyg.2012.00155
11. Bus AG, Leseman PP, Neuman SB (2012) Methods for preventing early academic difficulties. In: Harris KR, Graham S, Urdan T, Bus AG, Major S, Swanson HL (eds) APA Educational Psychology Handbook, vol 3, APA, pp 227–250
12. Chien NC, Howes C, Burchinal M, Pianta RC, Ritchie S, Bryant DM, Clifford RM, Early DM, Barbarin OA (2010) Childrens classroom engagement and school readiness gains in prekindergarten. Child Dev 81(5):1534–1549
13. Cicchetti DV (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 6(4):284
14. Clark HH (1996) Using language. Cambridge University Press, Cambridge
15. Denham SA, Blair KA, DeMulder E, Levitas J, Sawyer K, Auerbach-Major S, Queenan P (2003) Preschool emotional competence: pathway to social competence? Child Dev 74(1):238–256
16. Dominey PF, Dodane C (2004) Indeterminacy in language acquisition: the role of child directed speech and joint attention. J Neurolinguist 17(2–3):121–145. https://doi.org/10.1016/S0911-6044(03)00056-3
17. de Haas M, Vogt P, Krahmer E (2016) Enhancing child–robot tutoring interactions with appropriate feedback. In: Proceedings of the long-term child–robot interaction workshop at RO-MAN 2016
18. de Haas M, Baxter P, de Jong C, Vogt P, Krahmer E (2017) Exploring different types of feedback in preschooler- and robot interaction. In: Companion proceedings of the 2017 ACM/IEEE international conference on HRI
19. Erber NP (1975) Auditory-visual perception of speech. J Speech Hear Disord 40(4):481–492. https://doi.org/10.1044/jshd.4004.481
20. European Commission (2012) Special Eurobarometer 386: Europeans and their languages
21. Fridin M (2014a) Kindergarten social assistive robot: first meeting and ethical issues. Comput Hum Behav 30:262–272
22. Fridin M (2014b) Storytelling by a kindergarten social assistive robot: a tool for constructive learning in preschool education. Comput Educ 70:53–64
23. Glenberg AM (2008) Embodiment for education. Handbook of cognitive science: an embodied approach, pp 355–372
24. Gordon G, Breazeal C, Engel S (2015) Can children catch curiosity from a social robot? In: Proceedings of the 10th ACM/IEEE international conference on human–robot interaction, ACM, pp 91–98

25. Hald LA, de Nooijer J, van Gog T, Bekkering H (2016) Optimizing word learning via links to perceptual and motoric experience. Educ Psychol Rev 28(3):495–522. https://doi.org/10.1007/s10648-015-9334-2
26. Herberg J, Feller S, Yengin I, Saerbeck M (2015) Robot Watchfulness Hinders Learning Performance. In: Proceedings of the 24th IEEE international symposium on robot and human interactive communication, pp 153–160
27. Hirsh-Pasek K, Adamson LB, Bakeman R, Owen MT, Golinkoff RM, Pace A, Yust PKS, Suma K (2015) The contribution of early communication quality to low-income childrens language success. Psychol Sci 26(7):1071–1083. https://doi.org/10.1177/0956797615581493
28. Hockema SA, Smith LB (2009) Learning your language, outside-in and inside-out. Linguistics 47(2):453–479
29. 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(1):4–14. https://doi.org/10.1037/a0027238
30. Hoff E, Core C, Place S, Rumiche R, Seor M, Parra M (2012) Dual language exposure and early bilingual development. J Child Lang 39(1):1–27. https://doi.org/10.1017/S0305000910000759
31. Hoff E, Rumiche R, Burridge A, Ribot KM, Welsh SN (2014) Expressive vocabulary development in children from bilingual and monolingual homes: a longitudinal study from two to four years. Early Child Res Q 29(4):433–444. https://doi.org/10.1016/j.ecresq.2014.04.012
32. Hurtado N, Marchman VA, Fernald A (2008) Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in spanish-learning children. Dev Sci 11(6):F31–F39. https://doi.org/10.1111/j.1467-7687.2008.00768.x
33. Kanda T, Hirano T, Eaton D, Ishiguro H (2004) Interactive robots as social partners and peer tutors for children: a field trial. J Hum Comput Interact 19(1):61–84
34. Käser T, Klingler S, Schwing AG, Gross M (2014) Beyond knowledge tracing: modeling skill topologies with bayesian networks. In: Proceedings of the 12th international conference on intelligent tutoring systems. Springer, Berlin, pp 188–198
35. Kennedy J, Baxter P, Belpaeme T (2015) The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In: Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction, ACM, pp 67–74
36. Kennedy J, Baxter P, Senft E, Belpaeme T (2016) Social robot tutoring for child second language learning. In: Proceedings of the 11th ACM/IEEE international conference on human–robot interaction, ACM, pp 67–74
37. Kennedy J, Lemaignan S, Montassier C, Lavalade P, Irfan B, Papadopoulos F, Senft E, Belpaeme T (2017) Child speech recognition in human–robot interaction: evaluations and recommendations. In: Proceedings of the 12th ACM/IEEE international conference on human–robot interaction, ACM, pp 82–90
38. Konishi H, Kanero J, Freeman MR, Golinkoff RM, Hirsh-Pasek K (2014) Six principles of language development: implications for second language learners. Dev Neuropsychol 39(5):404–420. https://doi.org/10.1080/87565641.2014.931961
39. Kuhl PK (2007) Is speech learning 'gated' by the social brain? Dev Sci 10(1):110–120. https://doi.org/10.1111/j.1467-7687.2007.00572.x
40. Kuhl PK (2010) Brain mechanisms in early language acquisition. Neuron 67(5):713–727
41. Laevers F et al (2012) Pointing the compass to wellbeing-why and with what kind of result? Every Child 18(3):26
42. Lee S, Noh H, Lee J, Lee K, Lee GG, Sagong S, Kim M (2011) On the effectiveness of robot-assisted language learning. ReCALL 23(01):25–58
43. Leitner S (1972) So lernt man lernen. Der Weg zum Erfolg, Herder
44. Leseman PP, Rollenberg L, Rispens J (2001) Playing and working in kindergarten: cognitive co-construction in two educational situations. Early Child Res Q 16(3):363–384
45. Leyzberg D, Spaulding S, Toneva M, Scassellati B (2012) The physical presence of a robot tutor increases cognitive learning gains. In: Proceedings of the 34th annual conference of the cognitive science society
46. Li J (2015) The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int J Hum Comput Stud 77:23–37
47. Marulis LM, Neuman SB (2010) The Effects of vocabulary intervention on young children's word learning: a meta-analysis. Rev Educ Res 80(3):300–335. https://doi.org/10.3102/0034654310377087
48. Mol SE, Bus AG, de Jong MT (2009) Interactive book reading in early education: a tool to stimulate print knowledge as well as oral language. Rev Educ Res 79(2):979–1007
49. Moriguchi Y, Kanda T, Ishiguro H, Shimada Y, Itakura S (2011) Can young children learn words from a robot? Interact Stud 12(1):107–118
50. Movellan JR, Eckhardt M, Virnes M, Rodriguez A (2009) Sociable robot improves toddler vocabulary skills. In: Proceedings of the 4th ACM/IEEE international conference on human–robot interaction, pp 307–308
51. Neuman SB, Newman EH, Dwyer J (2011) Educational effects of a vocabulary intervention on preschoolers' word knowledge and conceptual development: a cluster-randomized trial. Read Res Q 46(3):249–272
52. Newcombe NS, Uttal DH, Sauter M (2013) Spatial development. Oxford handbook of developmental psychology 1:564–590
53. Oudgenoeg-Paz O, Verhagen J, Vlaar R (2017) Lesson series for three domains. Tech. rep., Universiteit Utrecht, L2TOR Deliverable
54. Pearson BZ, Fernndez SC, Oller DK (1993) Lexical development in bilingual infants and toddlers: comparison to monolingual norms. Lang Learn 43(1):93–120. https://doi.org/10.1111/j.1467-1770.1993.tb00174.x
55. Pearson BZ, Fernandez SC, Lewedeg V, Oller DK (1997) The relation of input factors to lexical learning by bilingual infants. Appl Psycholinguist 18(1):41–58. https://doi.org/10.1017/S0142716400009863
56. Place S, Hoff E (2011) Properties of dual language exposure that influence two-year-olds' bilingual proficiency. Child Dev 82(6):1834–1849. https://doi.org/10.1111/j.1467-8624.2011.01660.x
57. Proctor CP, Carlo M, August D, Snow C (2005) Native Spanish-speaking children reading in english: toward a model of comprehension. J Educ Psychol 97(2):246–256. https://doi.org/10.1037/0022-0663.97.2.246
58. Rosenthal-von der Pütten AM, Straßmann C, Krämer NC (2016) Robots or agents–neither helps you more or less during second language acquisition. In: Proceedings of the international conference on intelligent virtual agents. Springer, Berlin, pp 256–268
59. Rowe ML, Silverman RD, Mullan BE (2013) The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. Contemp Educ Psychol 38(2):109–117
60. Saerbeck M, Schut T, Bartneck C, Janse MD (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, ACM, pp 1613–1622. https://doi.org/10.1145/1753326.1753567
61. Schodde T, Bergmann K, Kopp S (2017) Adaptive robot language tutoring based on Bayesian knowledge tracing and predictive decision-making. In: Proceedings of the 12th ACM/IEEE international conference on human–robot interaction, ACM
62. Sidner CL, Lee C, Kidd CD, Lesh N, Rich C (2005) Explorations in engagement for humans and robots. Artif Intell 166(1–2):140–164

63. Spaulding S, Gordon G, Breazeal C (2016) Affect-aware student models for robot tutors. In: Proceedings of the 2016 international conference on autonomous agents and multiagent systems, pp 864–872
64. Tanaka F, Matsuzoe S (2012) Children teach a care-receiving robot to promote their learning: field experiments in a classroom for vocabulary learning. J Hum Robot Interact 1(1):78–95
65. Tellier M (2008) The effect of gestures on second language memorisation by young children. Gestures Lang Dev 8(2):219–235. https://doi.org/10.1075/gest.8.2.06tel
66. Tomasselo M, Todd J (1983) Joint attention and lexical acquisition style. First Lang 4:197–212
67. Unsworth S (2014) Comparing the role of input in bilingual acquisition across domains. In: Grter T, Paradis J (eds) Trends language acquisition research, vol 13, John Benjamins Publishing Company, pp 181–201. https://doi.org/10.1075/tilar.13.10uns
68. Vlaar R, Verhagen J, Oudgenoeg-Paz O, Leseman P (2017) Comparing L2 word learning through a tablet or real objects: what benefits learning most? In: Proceedings of the R4L workshop at HRI'17
69. Vogt P, de Haas M, de Jong C, Baxter P, Krahmer E (2017) Child-robot interactions for second language tutoring to preschool children. Frontiers in Human Neuroscience 11: https://doi.org/10.3389/fnhum.2017.00073
70. Vygotsky L (1978) Mind in society. Harvard University Press, Cambridge
71. Westlund JK, Dickens L, Jeong S, Harris P, DeSteno D, Breazeal C (2015) A comparison of children learning new words from robots, tablets, and people. In: Proceedings of the 1st international conference on social robots in therapy and education
72. Westlund KJM, Martinez M, Archie M, Das M, Breazeal C (2016a) Effects of framing a robot as a social agent or as a machine on children's social behavior. In: Proceedings of the 25th IEEE international symposium on robot and human interactive communication, IEEE, pp 688–693
73. Westlund JMK, Martinez M, Archie M, Das M, Breazeal C (2016b) A study to measure the effect of framing a robot as a social agent or as a machine on children's social behavior. In: The eleventh ACM/IEEE international conference on human robot interaction, IEEE Press, pp 459–460
74. Wermelinger S, Gampe A, Daum MM (2017) Bilingual toddlers have advanced abilities to repair communication failure. J Exp Child Psychol 155:84–94. https://doi.org/10.1016/j.jecp.2016.11.005
75. White J (1996) An input enhancement study with ESL children: effects on the acquisition of possessive determiners. Ph.D. dissertation, McGill University, Montreal, Canada
76. Yow WQ, Markman EM (2011) Young bilingual children's heightened sensitivity to referential cues. J Cognit Dev 12(1):12–31. https://doi.org/10.1080/15248372.2011.539524

Tony Belpaeme is a Professor at the Centre for Robotics and Neural Systems at the University of Plymouth (UK) and at Ghent University (Belgium). He received his Ph.D. in Computer Science from the Vrije Universiteit Brussel (Belgium). He leads a team studying cognitive robotics and human–robot interaction. Starting from the premise that intelligence is rooted in social interaction, Belpaeme and his research team try to further the science and technology behind artificial intelligence and social robots. This results in a spectrum of results, from theoretical insights to practical applications. He is the coordinator of the H2020 L2TOR project, a large-scale European project bringing 7 partners together to study how robots can be used to support the learning of a second language by children.

Paul Vogt is an Associate Professor at the Department of Cognitive Science and Artificial Intelligence at Tilburg University in the Netherlands. He received an M.Sc. in Cognitive Science and Engineering from the University of Groningen (Netherlands), and obtained a Ph.D. at the Artificial Intelligence Laboratory of the Vrije Universiteit Brussel (Belgium). His research focuses on understanding the cultural, social and cognitive mechanisms that underlie the evolution and acquisition of language and communication. Vogt is particularly interested in investigating how humans and machines can ground the meaning of linguistic utterances in the real world, and how they learn language from each other through social interactions. To study this, he has used a variety of techniques, ranging from agent-based modelling, child–robot interaction and psycholinguistic experiments to ethnographic research of children's language acquisition in different cultures.

Rianne van den Berghe is a Ph.D. candidate at the Department of Special Education at Utrecht University in the Netherlands. She received an MA in Linguistics from Utrecht University, in which she focused on (second) language acquisition and discourse processing. In her Ph.D. research, she investigates the way robot-assisted language lessons should be designed in order to optimize children's learning gains and engagement.

Kirsten Bergmann is a postdoctoral researcher at the Cluster of Excellence for Cognitive Interaction Technology at Bielefeld University in Germany. She received her Ph.D. in Computer Science from Bielefeld University. For the past ten years she has worked on empirically grounded and cognitively plausible models of multimodal communicative behaviour, with a particular focus on coordination of speech and gestures in artificial agents. In current projects she is developing embodied tutoring systems, and has been investigating the role of multimodal communication in educational settings.

Tilbe Göksun is an Assistant Professor of Psychology at the Department of Psychology at Koç University in Istanbul, Turkey. She received her Ph.D. in Developmental Psychology from Temple University in Philadelphia, PA, USA and worked as a postdoctoral researcher in the Center for Cognitive Neuroscience at the University of Pennsylvania in Philadelphia, PA, USA. She leads the Language and Cognition Lab at Koç University. Her research examines the interaction between language and thought processes, focusing on first and second language acquisition, event perception, relational and spatial language, neuropsychology of language, and the role of gestures in these processes.

Mirjam de Haas is a Ph.D. candidate at the Tilburg center for Cognition and Communication at Tilburg University in the Netherlands. She received her M.Sc. in Artificial Intelligence from the Radboud University Nijmegen, the Netherlands. Her Ph.D. focuses on the design of child robot tutor interactions and she is interested in how to keep the children motivated and engaged throughout the different lessons.

Junko Kanero is a Postdoctoral Researcher in the Department of Psychology at Koç University in Istanbul, Turkey. She received her Ph.D. in Developmental Psychology and Neuroscience from Temple University in Philadelphia, PA, USA. Her research examines language as a window into human cognition, using a broad range of methodologies including behavioural measures, eye tracking, EEG, and fMRI. She is interested in language development in infancy and childhood, neural

processing of language, and how language and non-linguistic cognition interact.

James Kennedy is a Postdoctoral Associate at Disney Research, Los Angeles, USA. James received his Ph.D. from Plymouth University, UK in 2017 for his work using social robots to tutor children. During his Ph.D., he worked as a Research Assistant on the EU-funded DREAM project and collaborated with the ALIZ-E, and L2TOR projects, focusing on the use of social robots in applications involving children. His research interests lie in Human–Robot Interaction and Socially Intelligent Agents.

Aylin C. Küntay is a professor of psychology and the Dean of College of Social Sciences and Humanities in Koç University. She received her Ph.D. in Developmental Psychology in 1997 from the University of California at Berkeley. Her work is on the relation of early communicative and language development to social interaction and cognitive development in bilingual and monolingual children.

Ora Oudgenoeg-Paz is a postdoctoral researcher at the Department of Special Education: Cognitive and Motor Disabilities at Utrecht University (the Netherlands). She received her Ph.D. in Pedagogics from Utrecht University. Her research focuses on early language and motor development and the link between the two. She studies how motor development, sensorimotor interactions and early language exposure facilitate the development of (spatial) cognition and (spatial) language. Her work concerns both first and second language acquisition.

Fotios Papadopoulos is a Reseach Fellow at Centre for Robotics and Neural Systems at the University of Plymouth (UK). He received his Ph.D. in 2012 from the University of Hertfordshire. His research interests lies within Human–Robot interaction, robot engagement, haptic communication, and robot tele-operation.

Thorsten Schodde is a Ph.D. candidate at the Cluster of Excellence for Cognitive Interaction Technology at Bielefeld University in Germany. He also received a Master of Science in Intelligent Systems from Bielefeld University. His Ph.D. research focuses on planning of an adaptive teaching course for second language learning lessons for preschool children. The major aim is the maximization of the learning gain while engagement and motivation are kept high.

Josje Verhagen is a post-doctoral researcher at the Department of Special Education: Cognitive and Motor Disabilities at Utrecht University, the Netherlands. She received her Ph.D. in 2009 on adult second language acquisition from the Max Planck in Psycholinguistics (Nijmegen) and Free University (Amsterdam). Her research interests are first and second language acquisition, and bilingualism. In her current research, she studies how specific properties of the language input affect acquisition, and how language development in young children relates to more general cognitive development.

Christopher D. Wallbridge is a Ph.D. candidate at the Centre for Robotics and Neural Systems at the University of Plymouth (UK). He received a Master of Science in Robotics from the University of Plymouth. His research is on the natural use of spatial concepts by robots, including their use in multiple languages.

Bram Willemsen is a Researcher at the Tilburg Center for Cognition and Communication at Tilburg University, the Netherlands. He received his M.Sc. in Communication and Information Sciences (cum laude) from Tilburg University, the Netherlands. His research interests concern problems related to Natural Language Understanding and data-driven approaches to dialogue modelling. His work within the L2TOR project focuses on the realization of context-aware generation of verbal and non-verbal robot behaviours.

Jan de Wit is a Ph.D. candidate at the Tilburg center for Cognition and Communication at Tilburg University in the Netherlands. He received his M.Sc. in Game and Media Technology from Utrecht University and his PDEng in User System Interaction from Eindhoven University of Technology, the Netherlands. He is on a quest to design technology that contributes to society in a fun and light-hearted way. In his Ph.D. research, he is exploring the role of robot-performed gestures in children's second language learning.

Vasfiye Geçkin is a lecturer at School of Foreign Languages, Bogazici University, Istanbul, Turkey. She received her Ph.D. (a cotutelle degree) in Linguistics from Macquarie University (Australia) and the University of Potsdam (Germany) in 2015, on acquisition of logical operators by monolingual and bilingual children. Her research focuses on language comprehension and production of bilingual and monolingual children.

Laura Hoffmann is Postdoctoral Researcher at the Cluster of Excellence Cognitive Interaction Technology (CITEC) associated with Bielefeld University, Germany. Her research is about the psychological impact of interactive technologies, in particular social robots. She uses quantitative and qualitative methods to understand how humans make sense of artificial others; how they perceive and evaluate them according to specific characteristics like embodiment, morphology or behavior.

Stefan Kopp is a Professor of Computer Science and head of the Social Cognitive Systems Group at Bielefeld University. He received his Ph.D. in 2004 for work on the synthesis of multimodal communicative behaviour of embodied agents. After a Postdoc at Northwestern University, he was research fellow at Bielefeld's Center for Interdisciplinary Research (ZiF) and is now principal investigator of the Center of Excellence "Cognitive Interaction Technology" (CITEC). Kopp and his team develop computational accounts of behavioural and cognitive abilities needed to act as a socially intelligent interaction partner. These are embedded in artificial systems like virtual 3D avatars or social robots, which are applied and evaluated in assisted living, industrial, or educational settings.

Emiel Krahmer is a Professor of Language, Cognition and Computation at the Tilburg School of Humanities and Digital Sciences, where he co-leads the Language, Communication and Cognition research group. He received his Ph.D. in Computational Linguistics in 1995, after which he worked as a postdoc in the Institute for Perception Research at the Eindhoven University of Technology before moving to Tilburg University. In his current work he studies how people communicate with each other, both in text and in speech, with the aim of subsequently improving the way computers communicate with human users. To achieve this, he combines computational modelling and experimental studies with human participants. Much of his

research is funded through external grants, including an NWO VICI grant.

Ezgi Mamus is a Ph.D. candidate at Radboud University in the Netherlands. She received her MA in Cognitive Psychology from Bogazici University in Istanbul, Turkey. Her Ph.D. research focuses on the influence of perceptual modality (e.g., sound vs. vision) on gestural representations of spatial events.

Jean-Marc Montanier is an engineer researcher at SoftBank Robotics Europe. He obtained his Ph.D. from Université Paris-Sud XI in 2013, on the study of autonomous auto-adaptation in swarm robotics. Since then he worked on topics relative to autonomous learning in multi-agent Systems. At SoftBank Robotics Europe, he is looking at the latests trends in Artificial Intelligence in order to transfer them to industrialized robots.

Cansu Oranç is a Ph.D. candidate at the Department of Psychology at Koç University in Istanbul, Turkey. She received her M.Sc. degree in Behavioural Science from Radboud University Nijmegen, the Netherlands. Her research focuses on the factors affecting children's learning of new information from different technological sources such as electronic books and robots.

Amit Kumar Pandey is Head Principal Scientist (Chief Scientist) at SoftBank Robotics Europe, Paris, France, also serving as the scientific coordinator and principal investigator of the European Union (EU) collaborative projects of the company. Earlier for 6 years he worked as researcher in Robotics and AI at LAAS-CNRS (French National Center for Scientific Research), France. Among other responsibilities, he is also the founding coordinator of Socially Intelligent Robots and Societal Applications (SIRo-SA) Topic Group (TG) of euRobotics. He is also the recipient of the second best Ph.D. thesis (tie) in Robotics in Europe for the prestigious euRobotics Georges Giralt Award.

Social Robots for Early Language Learning: Current Evidence and Future Directions

Junko Kanero,[1] Vasfiye Geçkin,[1,2] Cansu Oranç,[1] Ezgi Mamus,[1] Aylin C. Küntay,[1] and Tilbe Göksun[1]

[1]Koç University and [2]Boğaziçi University

ABSTRACT—***In this article, we review research on child–robot interaction (CRI) to discuss how social robots can be used to scaffold language learning in young children. First we provide reasons why robots can be useful for teaching first and second languages to children. Then we review studies on CRI that used robots to help children learn vocabulary and produce language. The studies vary in first and second languages and demographics of the learners (typically developing children and children with hearing and communication impairments). We conclude that, although social robots are useful for teaching language to children, evidence suggests that robots are not as effective as human teachers. However, this conclusion is not definitive because robots that tutor students in language have not been evaluated rigorously and technology is advancing rapidly. We suggest that CRI offers an opportunity for research and list possible directions for that work.***

KEYWORDS—*child–robot interaction; social robots; language learning*

Junko Kanero, Koç University; Vasfiye Geçkin, Koç University and Boğaziçi University; Cansu Oranç, Ezgi Mamus, Aylin C. Küntay, Tilbe Göksun, Koç University.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 688014.

Correspondence concerning this article should be addressed to Junko Kanero, Department of Psychology, Koç University, Rumelifeneri Yolu Sariyer, Istanbul, Turkey 34450; e-mail: jkanero@ku.edu.tr.

DOI: 10.1111/cdep.12277

Using technology in early education has gained considerable attention as digital devices (e.g., smartphones and tablets) have developed and been integrated into children's lives (1). In this article, we spotlight one of the newest additions to the list of devices—social robots—and discuss whether research on child–robot interaction (CRI) can help children learn first and second languages.

A social robot is "an autonomous or semiautonomous robot that interacts and communicates with humans by following the behavioral norms expected by the people with whom the robot is intended to interact" (2, p. 592). Social robots have been used to teach scientific knowledge, mathematics, social skills, computer programming, and language (3, 4). However, research on CRI has not been readily accessible to all those interested because the studies appear primarily in conference proceedings and journals dedicated to the field of robotics. Furthermore, these studies often focus on technical features of robots rather than educational concerns, such as whether and how robots can help young language learners.

In this article, we summarize findings on CRI and evaluate them critically. First we discuss briefly why a robot may be useful for teaching language to children. Then we evaluate whether children enjoy learning language with a robot. In the main section of the article, we ask whether children can learn language from a robot. We analyze learning outcomes at three levels: whether robots are *at all* successful teaching language to children, whether they are more successful teaching language than other digital devices, and whether robots can teach language as effectively as human teachers. Although social robots have potential as a language teaching tool, evidence suggests that robots are not as effective as human teachers. However, we argue that researchers must continue exploring this issue because the educational benefits of robots have not been evaluated thoroughly and technology in robotics is advancing quickly. In the last section, we suggest directions for research on CRI.

Child Development Perspectives

WHY USE ROBOTS FOR LEARNING LANGUAGE?

Learning language with a human teacher benefits children, but successful learning often takes more than just classes at school. Social robots are theorized to contribute to the early language learning experience in unique ways, and to supplement and enhance the experience. As a social agent with a physical body, a robot can play the role of a human through vocal, gestural, and facial expressions (5, 6). Although it remains unclear whether all the pedagogical strategies used by human teachers can and should be adopted by robots, many can be applied to robot-assisted language lessons (7).

One strength of robot tutors is their ability to perform actions and gestures. For example, a humanoid (a robot that resembles a human in appearance) can point to a physical object or open its arms to represent the meaning of the word *big*. Gestures abound in natural communication and can be a powerful cue that supplements speech. Robot tutors that can gesture may be especially effective for children because children benefit from gestures more than adults in human–human interaction (8); gestures improve speech comprehension in a second language (L2) in less-skilled learners (9), and gestures increase children's attention to the learning materials (10). For example, Italian-speaking 5- to 6-year-olds recalled stories more accurately when the tales were narrated by an expressive humanoid robot that used gestures, eye gaze, and voice tone than when they were told by an inexpressive human teacher (11).

Another strength of robots is that they are adaptive—through sensors, they can detect humans' motivational and educational needs and change their behavior accordingly. As suggested by scaffolding, learning outcomes are maximized when a task is not too difficult but challenging enough for a child (12). In one study, English-speaking 3- to 5-year-olds learned Spanish words successfully with a robot that provided explicit verbal feedback (e.g., "Good job!") as well as implicit feedback via eye gaze, a feature children often rely on in learning words (13), and adjusted them based on the children's performance (14). It can be difficult for classroom teachers to adjust lesson levels to each child and robot tutors can serve as a supplementary tool, especially when children can practice one on one with the robot.

In theory, social robots could provide unique support for young language learners. Does research confirm the idea? Next, we review empirical findings and evaluate whether children enjoy learning with a robot (in terms of motivation and engagement) and whether they can learn from a robot (in terms of learning outcomes).

MOTIVATION AND ENGAGEMENT

Motivation and engagement are popular measures in research on CRI. To understand whether children enjoy learning with a robot, studies of these factors have used children's self-reports to measure attention, satisfaction, and enjoyment (15, 16), and they have analyzed children's facial expressions (14). Although parents and educators may put less focus on engagement than on learning outcomes, engagement is a critical measure because children learn best when they are engaged (13). For robot-assisted lessons to be successful, children must want to continue to interact with robots.

Most children find learning language with social robots engaging (5, 14–21). For example, fifth graders in Taiwan practiced English skills in a group lesson led by a human teacher with or without a humanoid robot. Children who studied with the robot reported that they were more motivated and satisfied with the learning materials, and were less anxious and had greater self-esteem than their counterparts who studied without the robot (15). In another study, 3- to 5-year-old English speakers enjoyed learning fruit names in French with a robot (17). And in another study, Japanese preschoolers who learned English words from a humanoid robot were engaged and imitated the robot's movements as instructed (18). Interviews with children also suggest that children like robots (19) and prefer to learn from a robot than a tablet or a human (20). Positive attitudes toward robots have been observed both in class (21) and at home (5). Researchers working with children with autism spectrum disorder (ASD) also suggest that children's interest in robots contributes to their learning (22). Furthermore, teachers found a robot useful after using it in class (23; see also 24).

In summary, children enjoy learning language with a robot. However, we must interpret these findings cautiously because the advantage of robots may be due to novelty. Compared to a tablet or human teacher, the appearance of a robot is usually novel to children and can easily grab their attention. In a study in Japan, although elementary schoolers were initially very interested in interacting with a robot English tutor, after 1 week children interacted less frequently (25). To determine the motivational benefits of robots, researchers should explore interaction between robots and children for an extended time.

Research on motivation and engagement favors the use of robots in early language education. Although we must further examine whether children's engagement lasts, studies on CRI generally agree that learning with a social robot is exciting for children. However, the picture differs for learning outcomes.

LEARNING OUTCOMES

Research on vocabulary learning and language production provides a good ground for discussing learning outcomes of robot-assisted language lessons. Research in both domains has identified positive learning outcomes in robot-assisted language lessons, but the impact of robot language tutors varies across studies.

Vocabulary Learning

Vocabulary learning may be the most common topic in the field of CRI (19, 20, 26, 27). Researchers seem to agree that a social

robot can teach new words to children successfully. In one study, English-speaking 15- to 23-month-olds learned words with a robot that had a built-in touchscreen (26). The same pattern is also apparent in L2 acquisition: 3- to 5-year-old English speakers learned Spanish words over eight play sessions in which they were engaged in a tablet-based learning activity with a robot (14). In another study, Japanese-speaking 3- to 6-year-olds learned English verbs by teaching the words to the humanoid robot. These children identified corresponding pictures more successfully for the verbs they taught the robot than for the verbs they learned from a human experimenter, both on the day of the experiment and 3 to 5 weeks later (21). Although it remains unclear whether their learning improved due to the presence of the robot or because children taught the words to another agent, the study demonstrated the unique role robots can play in vocabulary learning.

Social robots may also help vocabulary development in children with ASD. Researchers in Iran developed a robot-assisted intervention to teach English words to Persian-speaking 7- to 9-year-olds with ASD. English test scores increased and were maintained after 2 weeks (28). However, in another study, after a 6-week intervention with a robot that involved imitation and games, English-speaking preschoolers with ASD and speech deficiency improved their receptive and expressive communication skills but did not improve their vocabulary (29).

Social robots have also helped children with hearing impairments. Researchers modified hands of a robot to sign Turkish Sign Language (TSL; 30). Six- to 16-year-old typically developing children and children with hearing impairments, as well as adults, understood and remembered TSL words generated by the robot and accurately matched the robot's sign gesture with the corresponding image. In another study by the same research group, 7- to 11-year-olds with beginner-level TSL skills learned more words when they interacted physically with the robot than when they watched the robot on a screen; 9- to 16-year-olds with advanced TSL skills learned equally well in both situations (31). The physical embodiment of robots may have different effects, depending on learners' language proficiency. The effectiveness of robots as sign language tutors has only been studied experimentally for TSL, though some have begun to examine their use in teaching other sign languages (e.g., Persian Sign Language; 32).

Young children can learn words from a robot. However, this does not necessarily mean that robots are more effective than other devices or humans in teaching language. In a 4-week reading program in Korean, 4-year-old native speakers learned stories either by interacting with a robot or by watching the stories on an electronic book. Children in both groups improved their vocabulary knowledge (27). In another study in which English-speaking 4- to 6-year-olds learned made-up words, children learned equally well from a robot, a human teacher, and a tablet (20). In yet another study of 4- to 6-year-olds, Italian-speaking children learned English words either with a robot or another child (33). And in a study with Japanese-speaking 4- and 5-year-olds, learning made-up words from a robot was not as effective as learning from a human (34).

To our knowledge, no study has found robots to be more effective at teaching words than other digital devices or human teachers, except for the sign language study in which beginners benefited from the physical presence of a robot (31). Sign language may be a promising direction because performing actions is a unique strength of robots. With regard to vocabulary learning, although further research may change the picture, robots may not confer more advantages than other mediums. However, the implications differ for language production.

Language Production

Social robots have been used to improve children's ability to produce language, for example, in storytelling skills (19, 27). In the study mentioned previously (27), Korean-speaking 4-year-olds learned vocabulary equally well with a robot and with an electronic book. However, only children who interacted with the robot improved their abilities to tell original stories, retell stories they learned, and recognize and pronounce written words. In addition, in another study, English-speaking 4- to 6-year-olds' own stories became longer and richer when the robot adjusted the lesson's complexity to children's language level (19).

Social robots can also elicit speech in children with ASD (22, 29). In the aforementioned 6-week intervention study, English-speaking preschoolers with ASD and speech deficiency produced more spontaneous speech after playing with a robot, although the study did not compare teaching by other devices or human teachers (29). Another study with English-speaking 4- to 12-year-olds with high-functioning ASD was more thorough (22): Children interacted in various combinations with adults, a touchscreen computer game, and a dinosaur robot. When interacting with the robot and an adult, children produced more utterances (toward the robot and the adult) than when they interacted with two adults or with the computer and one adult. These results suggest that, for children with ASD, a robot can be a more effective learning companion than computers or human adults.

For language production, some studies have demonstrated the benefit of robot companions over other digital devices. Social robots may be especially beneficial for individuals with ASD who face communication difficulties because practicing communication can be less intimidating with a robot than with another person (28, 35). We suggest that using robots in fostering language production is an important direction for research. Now, we turn to other research topics that should be explored.

LOOKING AHEAD

Research demonstrates that children are motivated to learn with a robot, but based on findings, we cannot claim that robot language tutors are particularly effective. Nonetheless, insufficient

evidence supporting the unique benefits of robot tutors should not be taken as definitive for two reasons: the dearth of empirical research and the advances of technology.

First, research may not have found robots to be more effective learning companions than other options because too few studies have been done. Studies on CRI are often descriptive and exploratory, and do not follow the scientific standards in other disciplines. Many lack a proper control group to evaluate whether a robot is more effective at teaching language than other options. Most studies have tested a small group of children and focused on whether children liked the robot, without evaluating learning outcomes. No research has examined long-term benefits of robot tutors. Furthermore, reports on CRI research often lack critical information (e.g., age of participants), making it difficult to evaluate the findings properly (36). Scholars in fields such as developmental psychology have examined language learning for decades, and incorporating their insights into designing and reporting experiments on CRI would be helpful, as would communicating with educators who use robots.

Second, we must consider advances in the hardware and software of robots. The technical features of robots that have been studied so far fail to meet the full potential of social robots, many of which may have completely different features within a few years. For example, developing a reliable system for recognizing children's speech automatically is a challenge because of factors such as the ungrammaticality of children's utterances and rapid developmental changes in the phonetic characteristics of children's speech. Currently available systems seem unreliable with children's speech, but different ways to improve the system have been suggested (37). When children's speech can be recognized reliably, robots can provide lessons that are more adaptive and interactive.

Furthermore, social robots may be more beneficial in teaching specific aspects of language (27) or specific groups of people (22). In addition to vocabulary learning and language production, other aspects of language (e.g., pronunciation) should also be explored (but see 16). Another topic worth investigating is the role of robots, which includes but is not limited to tutor (14), care receiver (21), and teaching assistant (15). Manipulating specific features of robots, such as adaptivity (19) and contingency (38), may also result in more effective learning. Because it is virtually impossible to draw a conclusion that applies to all robots, researchers should ask not whether robots are useful for teaching language but how robot language tutors can be improved.

Although we have a long way to go in researching CRI, some promising attempts have been made. L2TOR is a multisite project that aims to develop an autonomous humanoid robot for teaching L2 vocabulary (English, Dutch, and German) to 5-year-olds in three countries (Turkey, the Netherlands, and Germany; 39, 40), and that considers important points discussed in this article. First, the robot tutor will be compared directly to a tablet. Second, the target words are math and spatial concepts, many of which have conventional gestures robots can perform. Finally, the robot tutor will be evaluated over several weeks to examine long-term benefits in learning. Unlike most studies, L2TOR involves not only roboticists, but also developmental psychologists and linguists. Research on robot-assisted language learning is still at an early stage; strong interdisciplinary collaboration can help advance the field.

CONCLUSION

We have provided a concise, critical review of research on using social robots in early language education. Research suggests that robots may supplement a need that cannot be met solely by human teachers. However, when considering whether robots can substitute for other devices or human teachers, no study indicates that they are more effective than humans—though robots can be more effective than other digital devices. The shortage of evidence supporting the unique benefits of social robots should be viewed as an opportunity for researchers. We hope this article encourages interdisciplinary collaboration among experts on this important topic.

REFERENCES

1. Kim, Y., & Smith, D. (2017). Pedagogical and technological augmentation of mobile learning for young children interactive learning environments. *Interactive Learning Environments*, *25*, 4–16. https://doi.org/10.1080/10494820.2015.1087411
2. Bartneck, C., & Forlizzi, J. (2004, September). A design-centered framework for social human–robot interaction. *Proceedings of the 2004 IEEE International Workshop on Robot and Human Interactive Communication* (pp. 591–594). https://doi.org/10.1109/roman.2004.1374827
3. Mubin, O., Stevens, C. J., Shahid, S., Mahmud, A. A., & Dong, J. J. (2013). A review of the applicability of robots in education. *Technology for Education and Learning*, *1*, 1–7. https://doi.org/10.2316/Journal.209.2013.1.209-0015
4. Toh, L. P. E., Causo, A., Tzuo, P. W., Chen, I. M., & Yeo, S. H. (2016). A review on the use of robots in education and young children. *Journal of Educational Technology & Society*, *19*, 148–163.
5. Han, J. H., Jo, M. H., Jones, V., & Jo, J. H. (2008). Comparative study on the educational use of home robots for children. *Journal of Information Processing Systems*, *4*, 159–168. https://doi.org/10.3745/JIPS.2008.4.4.159
6. Kennedy, J., Baxter, P., & Belpaeme, T. (2015). Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics*, *7*, 293–308. https://doi.org/10.1007/s12369-014-0277-4
7. Vogt, P., de Haas, M., de Jong, C., Baxter, P., & Krahmer, E. (2017). Child–robot interactions for second language tutoring to preschool children. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00073
8. Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin*, *137*, 297–315. https://doi.org/10.1037/a0022128
9. Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language*

Learning, *55*, 661–699. https://doi.org/10.1111/j.0023-8333.2005.00320.x

10. Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, *28*, 187–204. https://doi.org/10.1016/S0361-476X(02)00007-3
11. Conti, D., Di Nuovo, A., Cirasa, C., & Di Nuovo, S. (2017). A comparison of kindergarten storytelling by human and humanoid robot with different social behavior. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human–Robot Interaction* (pp. 97–98), March 6–9, 2017, Vienna, Austria. https://doi.org/10.1145/3029798.3038359
12. Vygotsky, L. (1978). Interaction between learning and development. In M. Gauvain & M. Cole (Eds.), *Readings on the development of children* (pp. 34–40). New York, NY: Scientific American Books.
13. Konishi, H., Kanero, J., Freeman, M. R., Golinkoff, R. M., & Hirsh-Pasek, K. (2014). Six principles of language development: Implications for second language learners. *Developmental Neuropsychology*, *39*, 404–420. https://doi.org/10.1080/87565641.2014.931961
14. Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., … Breazeal, C. (2016). Affective personalization of a social robot tutor for children's second language skills. *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (pp. 3951–3957), February 12–17, 2016, Phoenix, AZ.
15. Hong, Z. W., Huang, Y. M., Hsu, M., & Shen, W. W. (2016). Authoring robot-assisted instructional materials for improving learning performance and motivation in EFL classrooms. *Educational Technology & Society*, *19*, 337–349.
16. Kim, J. W., & Kim, J. K. (2011). The effectiveness of robot pronunciation training for second language acquisition by children: Segmental and suprasegmental feature analysis approaches. *International Journal of Robots, Education and Art*, *1*, 1–17.
17. Freed, N. A. (2012). "This is the fluffy robot that only speaks French": Language use between preschoolers, their families, and a social robot while sharing virtual toys. Unpublished master's thesis, Massachusetts Institute of Technology, Cambridge, MA.
18. Tanaka, F., Isshiki, K., Takahashi, F., Uekusa, M., Sei, R., & Hayashi, K. (2015). Pepper learns together with children: Development of an educational application. *Proceedings of IEEE-RAS 15th International Conference on Humanoid Robots* (pp. 270–275), November 3–5, 2015, Seoul, South Korea. https://doi.org/10.1109/humanoids.2015.7363546
19. Kory Westlund, J. K., & Breazeal, C. (2015). The interplay of robot language level with children's language learning during storytelling. *Proceedings of the 10th Annual ACM/IEEE International Conference on Human–Robot Interaction* (pp. 65–66), March 2–5, 2015, Portland, OR. https://doi.org/10.1145/2701973.2701989
20. Kory Westlund, J. K., Dickens, L., Jeong, S., Harris, P., DeSteno, D., & Breazeal, C. (2015). A comparison of children learning new words from robots, tablets, and people. *New Friends: The 1st International Conference on Social Robots in Therapy and Education* (pp. 26–27), October 22–23, 2015, Almere, The Netherlands.
21. Tanaka, F., & Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human–Robot Interaction*, *1*, 78–95. https://doi.org/10.5898/JHRI.1.1.Tanaka
22. Kim, E. S., Berkovits, L. D., Bernier, E. P., Leyzberg, D., Shic, F., Paul, R., & Scassellati, B. (2013). Social robots as embedded reinforcers of social behavior in children with autism. *Journal of Autism and Developmental Disorders*, *43*, 1038–1049. https://doi.org/10.1007/s10803-012-1645-2
23. Chang, C. W., Lee, J. H., Chao, P. Y., Wang, C. Y., & Chen, G. D. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Journal of Educational Technology & Society*, *13*, 13–24.
24. Kory Westlund, J., Gordon, G., Spaulding, S., Lee, J. J., Plummer, L., Martinez, M., … Breazeal, C. (2016). Lessons from teachers on performing HRI studies with young children in schools. *Proceedings of the 11th ACM/IEEE International Conference on Human–Robot Interaction* (pp. 383–390), March 7–10, 2016, Christchurch, NZ.
25. Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction*, *19*, 61–84. https://doi.org/10.1207/s15327051hci1901&2_4
26. Movellan, J., Eckhardt, M., Virnes, M., & Rodriguez, A. (2009). Sociable robot improves toddler vocabulary skills. *Proceedings of the 4th ACM/IEEE International Conference on Human–Robot Interaction* (pp. 307–308), March 9–13, 2009, La Jolla, CA. https://doi.org/10.1145/1514095.1514189
27. Hyun, E. J., Kim, S. Y., Jang, S., & Park, S. (2008). Comparative study of effects of language instruction program using intelligence robot and multimedia on linguistic ability of young children. *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 187–192), August 1–3, 2008, Munich, Germany. https://doi.org/10.1109/roman.2008.4600664
28. Alemi, M., Meghdari, A., Basiri, N. M., & Taheri, A. (2015). The effect of applying humanoid robots as teacher assistants to help Iranian autistic pupils learn English as a foreign language. *Proceedings of the 7th International Conference on Social Robotics* (pp. 1–10), October 26–30, 2015, Paris, France.
29. Boccanfuso, L., Scarborough, S., Abramson, R. K., Hall, A. V., Wright, H. H., & 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. *Autonomous Robots*, *41*, 637–655. https://doi.org/10.1007/s10514-016-9554-4
30. Uluer, P., Akalın, N., & Köse, H. (2015). A new robotic platform for sign language tutoring. *International Journal of Social Robotics*, *7*, 571–585. https://doi.org/10.1007/s12369-015-0307-x
31. Köse, H., Uluer, P., Akalın, N., Yorgancı, R., Özkul, A., & Ince, G. (2015). The effect of embodiment in sign language tutoring with assistive humanoid robots. *International Journal of Social Robotics*, *7*, 537–548. https://doi.org/10.1007/s12369-015-0311-1
32. Zakipour, M., Meghdari, A., & Alemi, M. (2016). RASA: A low-cost upper-torso social robot acting as a sign language teaching assistant. *The 8th International Conference on Social Robotics* (pp. 630–639), November 1–3, 2016, Kansas City, MO. https://doi.org/10.1007/978-3-319-47437-3_62
33. Mazzoni, E., & Benvenuti, M. (2015). A robot-partner for preschool children learning English using socio-cognitive conflict. *Journal of Educational Technology & Society*, *18*, 474–485.
34. Moriguchi, Y., Kanda, T., Ishiguro, H., Shimada, Y., & Itakura, S. (2011). Can young children learn words from a robot? *Interaction Studies*, *12*, 107–118. https://doi.org/10.1075/is.12.1.04mor
35. Kozima, H., Nakagawa, C., & Yasuda, Y. (2005). Interactive robots for communication-care: A case-study in autism therapy. *IEEE International Workshop on Robots and Human Interactive Communication* (pp. 341–346). August 13–15, 2005, Nashville, TN.

36. Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., & Belpaeme, T. (2016). From characterising three years of HRI to methodology and reporting recommendations. *The 11th ACM/IEEE International Conference on Human Robot Interaction* (pp. 391–398), March 7–10, 2016, Christchurch, NZ.
37. Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., … Belpaeme, T. (2017). Child speech recognition in human–robot interaction: Evaluations and recommendations. *Proceedings of the 12th Annual ACM/IEEE International Conference on Human–Robot Interaction* (pp. 82–90), March 6–9, 1017, Vienna, Austria. https://doi.org/10.1145/2909824.302022937
38. Park, H. W., Gelsomini, M., Lee, J. J., & Breazeal, C. (2017). Telling stories to robots: The effect of backchanneling on a child's storytelling. *Proceedings of the 12th Annual ACM/IEEE International Conference on Human–Robot Interaction* (pp. 100–108), March 6–9, 2017, Vienna, Austria, https://doi.org/10.1145/2909824.3020245
39. Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. E. J., Kopp, S., … Deblieck, T. (2015). L2TOR—Second language tutoring using social robots. *Proceedings of 1st International Workshop on Educational Robotics* (pp. 100–108), October 26, 2015, Paris, France.
40. Belpaeme, T., Vogt, P., van den Berghe, R., Bergmann, K., Göksun, T., … Pandey, A. K. (2018). Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*.

HUMAN-ROBOT INTERACTION

Social robots for education: A review

Tony Belpaeme[1,2]*, James Kennedy[2], Aditi Ramachandran[3], Brian Scassellati[3], Fumihide Tanaka[4]

Social robots can be used in education as tutors or peer learners. They have been shown to be effective at increasing cognitive and affective outcomes and have achieved outcomes similar to those of human tutoring on restricted tasks. This is largely because of their physical presence, which traditional learning technologies lack. We review the potential of social robots in education, discuss the technical challenges, and consider how the robot's appearance and behavior affect learning outcomes.

INTRODUCTION

Virtual pedagogical agents and intelligent tutoring systems (ITSs) have been used for many years to deliver education, with comprehensive reviews available for each field (*1*, *2*). The use of social robots has recently been explored in the educational domain, with the expectation of similarly positive benefits for learners (*3*–*5*). A recent survey of long-term human-robot interaction (HRI) highlighted the increasing popularity of using social robots in educational environments (*6*), and restricted surveys have previously been conducted in this domain (*7*, *8*).

In this paper, we present a review of social robots used in education. The scope was limited to robots that were intended to deliver the learning experience through social interaction with learners, as opposed to robots that were used as pedagogical tools for science, technology, engineering, and math (STEM) education. We identified three key research questions: How effective are robot tutors at achieving learning outcomes? What is the contribution made by the robot's appearance and behavior? And what are the potential roles of a robot in an educational setting? We support our review with data gleaned from a statistical meta-analysis of published literature. We aim to provide a platform for researchers to build on by highlighting the expected outcomes of using robots to deliver education and by suggesting directions for future research.

Benefits of social robots as tutoring agents

The need for technological support in education is driven by demographic and economic factors. Shrinking school budgets, growing numbers of students per classroom, and the demand for greater personalization of curricula for children with diverse needs are fueling research into technology-based support that augments the efforts of parents and teachers. Most commonly, these systems take the form of a software system that provides one-on-one tutoring support. Social interaction enhances learning between humans, in terms of both cognitive and affective outcomes (*9*, *10*). Research has suggested that some of these behavioral influences also translate to interactions between robots and humans (*3*, *11*). Although robots that do not exhibit social behavior can be used as educational tools to teach students about technology [such as in (*12*)], we limited our review to robots designed specifically to support education through social interactions.

Because virtual agents (presented on laptops, tablets, or phones) can offer some of the same capabilities but without the expense of additional hardware, the need for maintenance, and the challenges of distribution and installation, the use of a robot in an educational setting must be explicitly justified. Compared with virtual agents, physically embodied robots offer three advantages: (i) they can be used for curricula or populations that require engagement with the physical world, (ii) users show more social behaviors that are beneficial for learning when engaging with a physically embodied system, and (iii) users show increased learning gains when interacting with physically embodied systems over virtual agents.

Robots are a natural choice when the material to be taught requires direct physical manipulation of the world. For example, tutoring physical skills, such as handwriting (*13*) or basketball free throws (*14*), may be more challenging with a virtual agent, and this approach is also taken in many rehabilitation- or therapy-focused applications (*15*). In addition, certain populations may require a physically embodied system. Robots have already been proposed to aid individuals with visual impairments (*16*) and for typically developing children under the age of two (*17*) who show only minimal learning gains when provided with educational content via screens (*18*).

In addition, often there is an expectation for robot tutors to be able to move through dynamic and populated spaces and manipulate the physical environment. Although not always needed in the context of education, there are some scenarios where the learning experience benefits from the robot being able to manipulate objects and move autonomously, such as when supporting physical experimentation (*19*) or moving to the learner rather than the learner moving to the robot. These challenges are not exclusive to social robotics and robot tutors, but the added elements of having the robot operate near and with (young) learners add complexities that are often disregarded in navigation and manipulation.

Physical robots are also more likely to elicit from users social behaviors that are beneficial to learning (*20*). Robots can be more engaging and enjoyable than a virtual agent in cooperative tasks (*21*–*23*) and are often perceived more positively (*22*, *24*, *25*). Importantly for tutoring systems, physically present robots yield significantly more compliance to its requests, even when those requests are challenging, than a video representation of the same robot (*26*).

Last, physical robots have enhanced learning and affected later behavioral choice more substantially than virtual agents. Compared with instructions from virtual characters, videos of robots, or audio-only lessons, robots have produced more rapid learning in cognitive puzzles (*27*). Similar results have been demonstrated when coaching users to select healthier snacks (*24*) and when helping users continue a 6-week weight-loss program (*28*). A comprehensive review (*25*) concluded that the physical presence of a robot led to positive perceptions

[1]Ghent University, Ghent, Belgium. [2]University of Plymouth, Plymouth, UK. [3]Yale University, New Haven, CT 06520–8285, USA. [4]University of Tsukuba, Tsukuba, Japan.
*Corresponding author. Email: tony.belpaeme@ugent.be

and increased task performance when compared with virtual agents or robots displayed on screens.

Technical challenges of building robot tutors

There are a number of challenges in using technology to support education. Using a social robot adds to this set of challenges because of the robot's presence in the social and physical environment and because of the expectations the robot creates in the user. The social element of the interaction is especially difficult to automate: Although robot tutors can operate autonomously in restricted contexts, fully autonomous social tutoring behavior in unconstrained environments remains elusive.

Perceiving the social world is a first step toward being able to act appropriately. Robot tutors should be able to not only correctly interpret the user's responses to the educational content offered but also interpret the rapid and nuanced social cues that indicate task engagement, confusion, and attention. Although automatic speech recognition and social signal processing have improved in recent years, sufficient progress has not been made for all populations. Speech recognition for younger users, for example, is still insufficiently robust for most interactions (*29*). Instead, alternative input technologies, such as a touch-screen tablets or wearable sensors, are used to read responses from the learner and can be used as a proxy to detect engagement and to track the performance of the student (*30*–*32*). Robots can also use explicit models of disengagement in a given context (*33*) and strategies, such as activity switching, to sustain engagement over the interaction (*34*). Computational vision has made great strides in recent years but is still limited when dealing with the range of environments and social expressions typically found in educational and domestic settings. Although advanced sensing technologies for reading gesture, posture, and gaze (*35*) have found their way into tutoring robots, most social robot tutors continue to be limited by the degree to which they can accurately interpret the learner's social behavior.

Armed with whatever social signals can be read from the student, the robot must choose an action that advances the long-term goals of the educational program. However, this can often be a difficult choice, even for experienced human instructors. Should the instructor press on and attempt another problem, advance to a more challenging problem, review how to solve the current problem, offer a hint, or even offer a brief break from instruction? There are often conflicting educational theories in human-based instruction, and whether or not these same theories hold when considering robot instructors is an open question. These choices are also present in ITSs, but the explicit agentic nature of robots often introduces additional options and, at times, complications. Choosing an appropriate emotional support strategy based on the affective state of the child (*36*), assisting with a meta-cognitive learning strategy (*37*), deciding when to take a break (*31*), and encouraging appropriate help-seeking behavior (*4*) have all been shown to increase student learning gains. Combining these actions with appropriate gestures (*38*), appropriate and congruent gaze behavior (*39*), expressive behaviors and attention-guiding behaviors (*11*), and timely nonverbal behaviors (*3*) also positively affects student recall and learning. However, merely increasing the amount of social behavior for a robot does not lead to increased learning gains: Certain studies have found that social behavior may be distracting (*40*, *41*). Instead, the social behavior of the robot must be carefully designed in conjunction with the interaction context and task at hand to enhance the educational interaction.

Last, substantial research has focused on personalizing interactions to the specific user. Within the ITS community, computational techniques such as dynamic Bayesian networks, fuzzy decision trees, and hidden Markov models are used to model student knowledge and learning. Similar to on-screen tutoring systems, robot tutors use these same techniques to help tailor the complexity of problems to the capabilities of the student, providing more complex problems only when easier problems have been mastered (*42*–*44*). In addition to the selection of personalized content, robotic tutoring systems often provide additional personalization to support individual learning styles and interaction preferences. Even straightforward forms of personalization, such as using a child's name or referencing personal details within an educational setting, can enhance user perception of the interaction and are important factors in maintaining engagement within learning interactions (*45*, *46*). Other affective personalization strategies have been explored to maintain engagement during a learning interaction by using reinforcement learning to select the robot's affective responses to the behavior of children (*47*). A field study showed that students who interacted with a robot that simultaneously demonstrated three types of personalization (nonverbal behavior, verbal behavior, and adaptive content progression) showed increased learning gains and sustained engagement when compared with students interacting with a nonpersonalized robot (*48*) Although progress has been made in constituent technologies of robot tutors—from perception to action selection and production of behaviors that promote learning—the integration of these technologies and balancing their use to elicit prosocial behavior and consistent learning still remain open challenges.

REVIEW

To support our review, we used a meta-analysis of the literature on robots for education. In this, three key questions framed the meta-analysis and dictated which information was extracted:

1. Efficacy. What are the cognitive and affective outcomes when robots are used in education?

2. Embodiment. What is the impact of using a physically embodied robot when compared with alternative technologies?

3. Interaction role. What are the different roles the robot can take in an educational context?

For the meta-analysis, we used published studies extracted from the Google Scholar, Microsoft Academic Search, and CiteSeerX databases by using the following search terms: robot tutor, robot tutors, socially assistive robotics (with manual filtering of those relevant to education), robot teacher, robot assisted language learning, and robot assisted learning. The earliest published work appeared in 1992, and the survey cutoff date was May 2017. In addition, proceedings of prominent social HRI journals and conferences were manually searched for relevant material: International Conference on Human-Robot Interaction, *International Journal of Social Robotics*, *Journal of Human-Robot Interaction*, International Conference on Social Robotics, and the International Symposium on Robot and Human Interactive Communication (RO-MAN).

The selection of papers was based on four additional criteria:

1) Novel experimental evaluations or analyses should be presented.

2) The robot should be used as the teacher (i.e., the robot is an agent in the interaction) rather than the robot being used as an educational prop or a learner with no intention to educate [e.g., (*49*)].

3) The work must have included a physical robot, with an educative intent. For example, studies considering "coaches" that sought to improve motivation and compliance, but did not engage in education [e.g., (*50*)], were not included, whereas those that provided tutoring and feedback were included [e.g., (*15*)].

4) Only full papers were included. Extended abstracts were omitted because these often contained preliminary findings, rather than complete results and full analyses.

We withheld 101 papers for analysis and excluded 12 papers for various reasons (e.g., the paper repeated results from an earlier publication). The analyzed papers together contain 309 study results (*51*).

To compare outcomes of the different studies, we first divided the outcomes of an intervention into either affective or cognitive. Cognitive outcomes focus on one or more of the following competencies: knowledge, comprehension, application, analysis, synthesis, and evaluation (*52*–*54*). Affective outcomes refer to qualities that are not learning outcomes per se, for example, the learner being attentive, receptive, responsive, reflective, or inquisitive (*53*). The meta-analysis contained 99 (33.6%) data points on cognitive learning outcomes and 196 (66.4%) data points on affective learning outcomes; 14 study results did not contain a comparative experiment on learning outcomes.

Cognitive outcomes are typically measured through pre- and posttests of student knowledge, whereas affective outcomes are more varied and can include self-reported measures and observations by the experimenters. Table 1 contains the most common methods for measuring cognitive and affective outcomes reported in the literature.

Most studies focused on children (179 data points; 58% of the sample; mean age, 8.2 years; SD, 3.56), whereas adults (≥18 years old) were a lesser focus of research in robot tutoring (98 data points; 32% of the sample; mean age, 30.5; SD, 17.5). For 29 studies (9%), both children and adults were used, or the age of the participants was not specified.

If the results reported an effect size expressed as Cohen's *d*, then this was used unaltered. In cases where the effect size was not reported or if it was expressed in a measure other than Cohen's *d*, then an online calculator (*55*) [see also (*56*)] was used if enough statistical information was present in the paper (typically participant numbers, means, and SDs are sufficient).

We captured the following data gleaned from the publications: the study design, the number of conditions, the number of participants per condition, whether participants were children or adults, participant ages (mean and SD), the robot used, the country in which the study was run, whether the study used a within or between design, the reported outcomes (affective or cognitive, with details on what was measured exactly), the descriptive statistics (where available mean, SD, *t*, and *F* values), the effect size as Cohen's *d*, whether the study involved one robot teaching one person or one robot teaching many, the role of the robot (presenter, teaching assistant, teacher, peer, or tutor), and the topic under study (embodiment of the robot, social character of the robot, the role of the robot, or other).

The studies in our sample reported more on affective outcomes than cognitive outcomes (Fig. 1A). This is due to the relative ease with which a range of affective outcomes can be assessed by using questionnaires and observational studies, whereas cognitive outcomes require administering a controlled knowledge assessment before and after the interaction with the robot, of which typically only one is reported per study.

Figure 2B shows the countries where studies were run. Robots for learning research, perhaps unsurprisingly, happen predominantly in East Asia (Japan, South Korea, and Taiwan), Europe, and the United States. An exception is the research in Iran on the use of robots to teach English in class settings.

Table 1. Common measures for determining cognitive and affective outcomes in robots for learning.

Cognitive	Learning gain, measured as difference between pre- and posttest score
	Administer posttest either immediately after exposure to robot or with delay
	Correct for varying initial knowledge, e.g., using normalized learning gain (*77*)
	Difference in completion time of test
	Number of attempts needed for correct response
Affective	Persistence, measured as number of attempts made or time spent with robot
	Number of interactions with the system, such as utterances or responses
	Coding emotional expressions of the learner, can be automated using face analysis software (*47*)
	Godspeed questionnaire, measuring the user's perception of robots (*78*)
	Tripod survey, measuring the learner's perspective on teaching, environment, and engagement (*79*)
	Immediacy, measuring psychological availability of the robot teacher (*3*, *10*)
	Evolution of time between answers, e.g., to indicate fatigue (*31*)
	Coding of video recordings of participants responses
	Coding or automated recording of eye gaze behavior (to code attention, for example)
	Subjective rating of the robot's teaching and the learning experience (*15*)
	Foreign language anxiety questionnaire (*80*)
	KindSAR interactivity index, quantitative measure of children's interactions with a robot (*81*)
	Basic empathy scale, self-report of empathy (*82*)
	Free-form feedback or interviews

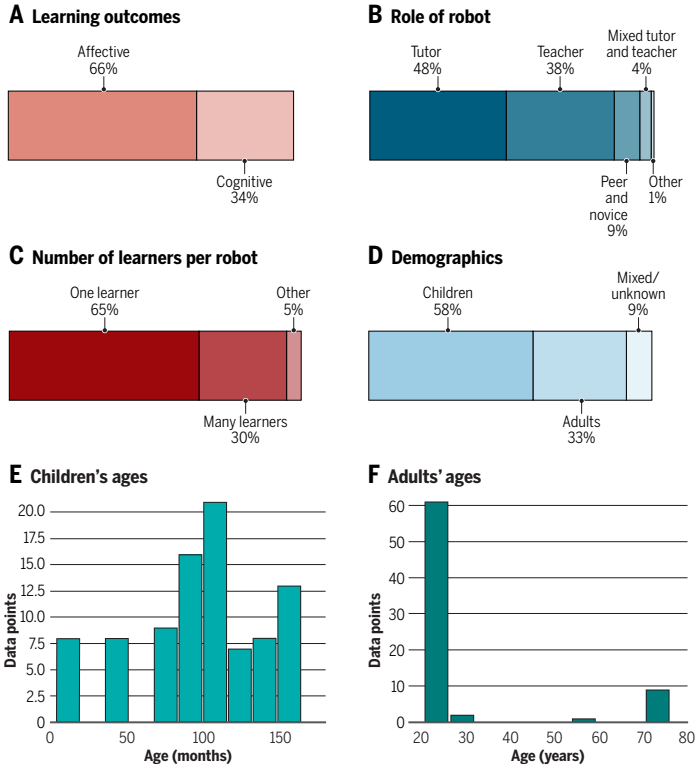


Fig. 1. An overview of data from the meta-analysis. (**A**) Type of learning outcome studied. (**B**) Role of the robot in the interaction. (**C**) Number of learners per robot in studies. (**D**) Division between children and adults (≥18 years old). (**E**) Age distribution for children. (**F**) Age distribution for adults.

Extracting meaningful statistical data from the published studies is not straightforward. Of the 309 results reported in 101 published studies, only 81 results contained enough data to calculate an effect size, highlighting the need for more rigorous reporting of data in HRI.

Efficacy of robots in education

The efficacy of robots in education is of primary interest, and here, we discuss the outcomes that might be expected when using a robot in education. The aim is to provide a high-level overview of the effect size that might be expected when comparing robots with a variety of control conditions, grouping a range of educational scenarios with many varying factors between studies (see Fig. 3). More specific analyses split by individual factors will be explored in subsequent sections.

Learning effects are divided into cognitive and affective outcomes. Across all studies included in the meta-review, we have 37 results that compared the robot with an alternative, such as an ITS, an on-screen avatar, or human tutoring. Of these, the aggregated mean cognitive outcome effect size (Cohen's d weighted by N) of robot tutoring is 0.70 [95% confidence interval (CI), 0.66 to 0.75] from 18 data points, with a mean of $N = 16.9$ participants per data point. The aggregated mean affective outcome effect size (Cohen's d weighted by N) is 0.59 (95% CI, 0.51 to 0.66) from 19 data points, with a mean of $N = 24.4$ students per data point. Many studies using robots do not consider learning in comparison with an alternative, such as computer-based or human tutoring, but instead against other versions of the same robot with different behaviors. The limited number of studies that did compare a robot against an alternative offers a positive picture of the contribution to learning made by social robots, with a medium effect size for affective and cognitive outcomes. Furthermore, positive affective outcomes did not imply positive cognitive outcomes, or vice versa. In some studies, introducing a robot improved affective outcomes while not necessarily leading to significant cognitive gains [e.g. (*57*)].

Human tutors provide a gold standard benchmark for tutoring interactions. Trained tutors are able to adapt to learner needs and modify strategies to maximize learning (*58*). Previous work (*59*) has suggested that human tutors produce a mean cognitive outcome effect size (Cohen's d) of 0.79, so the results observed when using a robot are in a similar region. However, social robots are typically deployed in restricted scenarios: short, well-defined lessons delivered with limited adaptation to individual learners or flexibility in curriculum. There is no suggestion yet that robots have the capability to tutor in a general sense as well as a human can. Comparisons between robots and humans are rare in the literature, so no meta-analysis data were available to compare the cognitive learning effect size.

Robot appearance

Because the positive learning outcomes are driven by the physical presence of the robot, the question remains of what exactly it is about the robot's appearance that promotes learning. A wide range of robots have been used in the surveyed studies, from small toy-like robots to full-sized android robots. Figure 2A shows the most used robots in the published studies.

The most popular robot in the studies we analyzed is the Nao robot, a 54-cm-tall humanoid by Softbank Robotics Europe available as having 14, 21, or 25 degrees of freedom (see Fig. 4B). The two latter versions of Nao have arms, legs, a torso, and a head. They can walk, gesture, and pan and tilt their head. Nao has a rich sensor suite and an on-board computational core, allowing the robot to be fully autonomous. The dominance of Nao for HRI can be attributed to its wide availability, appealing appearance, accessible price point, technical robustness, and ease of programming. Hence, Nao has become an almost de facto platform for many studies in robots for learning. Another robot popular as a tutor is the Keepon robot, a consumer-grade version of the Keepon Pro research robot. Keepon is a 25-cm-tall snowman-shaped robot with a yellow foam exterior without arms and legs (see Fig. 4C). It has four degrees of freedom to make it pan, roll, tilt, and bop. Originally sold as a novelty for children, it can be used as a research platform after some modification. Nao and Keepon offer two extremes in the design space of social robots, and hence, it is interesting to compare learning outcomes for both.

Comparing Keepon with Nao, the respective cognitive learning gain is $d = 0.56$ ($N = 10$; 95% CI, 0.532 to 0.58) and $d = 0.76$ ($N = 8$; 95% CI, 0.52 to 1.01); therefore, both show a medium-sized effect. However, we note that direct comparisons between different robots are difficult with the available data, because no studies used the same experimental design, the same curriculum, and the same student population with multiple robots. Furthermore, different robots have tended to be used at different times, becoming popular in studies when that particular hardware model was first made available and decreasing in usage over time. Because the complexity of the experimental protocols has tended to increase, direct comparison is not possible at this point in time.

What is clear from surveying the different robot types is that all robots have a distinctly social character [except for the Heathkit

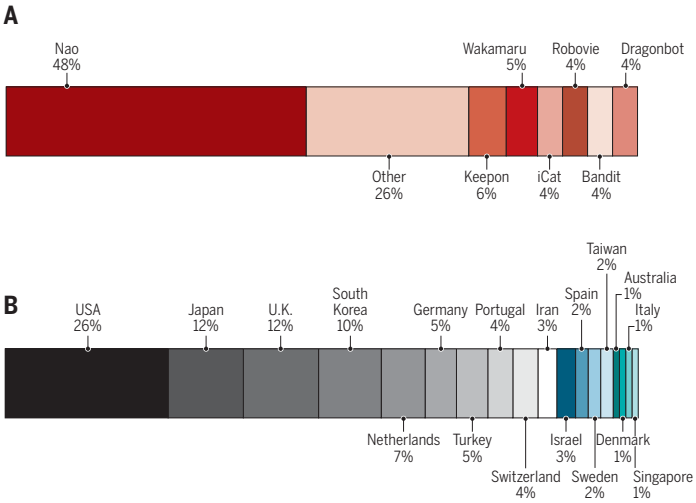


Fig. 2. Diversity of robots in education. (**A**) Types of robots used in the studies. (**B**) Nations where the research studies were run.

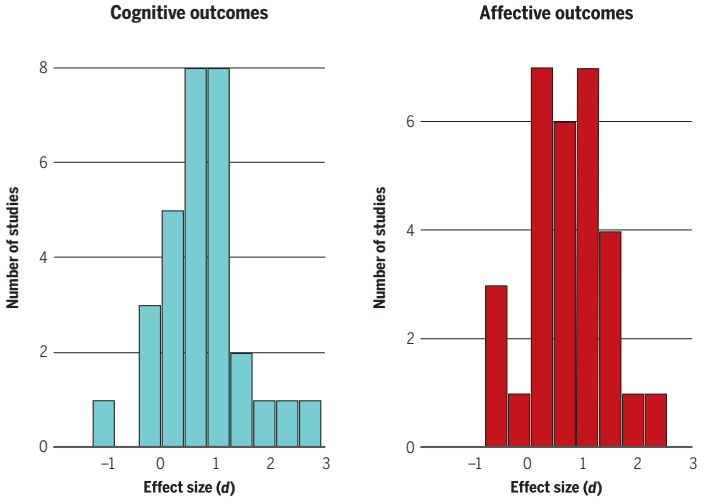


Fig. 3. Histograms of effect sizes (Cohen's *d*) for all cognitive and affective outcomes of robot tutors in the meta-analysis. These combine comparisons between robots and alternative educational technologies but also comparisons between different implementations of the robot and its tutoring behavior. In the large majority of results, adding a robot or adding supportive behavior to the robot improves outcomes.

HERO robot used in (*60*)]. All robots have humanoid features—such as a head, eyes, a mouth, arms, or legs—setting the expectation that the robot has the ability to engage on a social level. Although there are no data on whether the social appearance of the robot is a requirement for effective tutoring, there is evidence that the social and agentic nature of the robots promotes secondary responses conducive to learning (*61*, *62*). The choice of robot very often depends on practical considerations and whether the learners feel comfortable around the robot. The weighted average height of the robots is 62 cm; the shortest robot in use is the Keepon at 25 cm, and the tallest is the RoboThespian humanoid at 175 cm. Shorter robots are often preferred when teaching young children.

Robot behavior

To be effective educational agents, the behavior of social robots must be tailored to support various aspects of learning across different learners and diverse educational contexts. Several studies focused on understanding critical aspects of educational interactions to which robots should respond, as well as determining both what behaviors social robots can use and when to deliver these behaviors to affect learning outcomes.

Our meta-review shows that almost any strategy or social behavior of the robot aimed at increasing learning outcomes has a positive effect. We identified the influence of robot behaviors on cognitive outcomes ($d = 0.69$; $N = 12$; 95% CI, 0.56 to 0.83) and affective outcomes ($d = 0.70$; $N = 32$; 95% CI, 0.62 to 0.77).

Similar to findings in the ITS community, robots that personalize what content to provide based on user performance during an interaction can increase cognitive learning gains (*43*, *44*). In addition to the adaptive delivery of learning material, social robots can offer socially supportive behaviors and personalized support for learners within an educational context. Personalized social support, such as using a child's name or referring to previous interactions (*45*, *46*), is the low-hanging fruit of social interaction. More complex prosocial behavior, such as attention-guiding (*11*), displaying congruent gaze behavior (*39*), nonverbal immediacy (*3*), or showing empathy with the learner (*36*), not only has a positive impact on affective outcomes but also results in increased learning.

However, just as human tutors must at times sit quietly and allow students the opportunity to concentrate on problem solving, robot tutors must also limit their social behavior at appropriate times based on the cognitive load and engagement of the student (*40*). The social behavior of the robot must be carefully designed in conjunction with the interaction context and task at hand to enhance the educational interaction and avoid student distraction.

It is possible that the positive cognitive and affective learning outcomes of robot tutors are not directly caused by the robot having a physical presence, but rather the physical presence of the robot promotes social behaviors in the learner that, in turn, foster learning and create a positive learning experience. Robots have been shown to have a positive impact on compliance (*26*), engagement (*21*–*23*), and conformity (*20*), which, in turn, are conducive to achieving learning gains. Hence, a perhaps valuable research direction is to explore what it is about social robots that affects the first-order outcomes of engagement, persuasion, and compliance.

Robot role

Social robots for education include a variety of robots having different roles. Beyond the typical role of a teacher or a tutor, robots can also support learning through peer-to-peer relationships and can support skill consolidation and mastery by acting as a novice. In this section, we provide an overview of the different roles a robot can adopt and what their educational benefits are.

Robot as tutor or teacher

As a tutor or teacher, robots provide direct curriculum support through hints, tutorials, and supervision. These types of educational robots, including teaching assistant robots (*63*), have the longest history of research and development, often targeting curricular domains for young children. Early field studies placed robots into classrooms to observe whether they would have any qualitative impact on the learners' attitude and progress, but current research tends toward controlled experimental trials in both laboratory settings and classrooms (*64*).

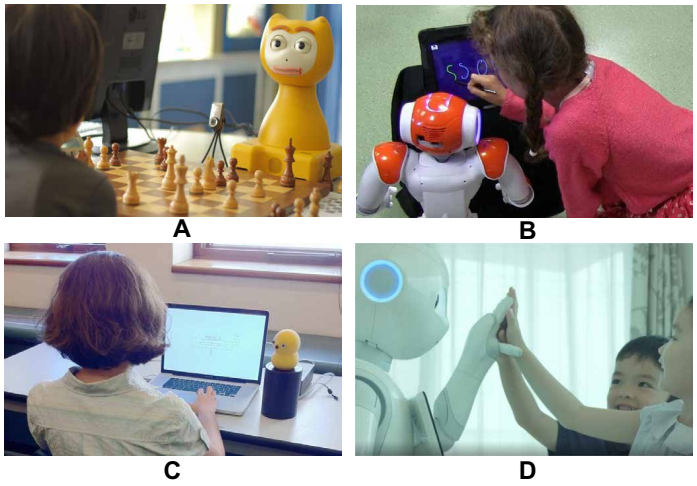


Fig. 4. Illustrative examples of social robots for learning. (**A**) iCat robot teaching young children to play chess (*76*). (**B**) Nao robot supporting a child to improve her handwriting (*13*). (**C**) Keepon robot tutoring an adult in a puzzle game (*27*). (**D**) Pepper robot providing motivation during English classes for Japanese children (*74*).

A commercial tutor robot called IROBI (Yujin Robotics) was released in the early 2000s. Designed to teach English, IROBI was shown to enhance both concentration on learning activities and academic performance compared with other teaching technology, such as audio material and a web-based application (*65*).

The focus on younger children links robot education research with other scientific areas, such as language development and developmental psychology (*66*). On the basis of the earlier work that studied socialization between toddlers and robots in a nursery school (*67*), a fully autonomous robot was deployed in classrooms. It was shown that the vocabulary skills of 18- to 24-month-old toddlers improved significantly (*68*). Much of the work in which the robot is used as a tutor focuses on one-to-one interactions, because these offer the greatest potential for personalized education.

In some cases, the robot is used as a novel channel through which a lecture is delivered. In these cases, the robot is not so much interacting with the learners but acts as a teacher or an assistant for the teacher (*69*). The value of the robot in this case lies in improving attention and motivation in the learners, while the delivery and assessment is done by the human teacher. Here, the delivery is often one to many, with the robot addressing an entire group of learners (*33*, *63*, *69*).

Robot as peer

Robots can also be peers or learning companions for humans. Not only does a peer have the potential of being less intimidating than a tutor or teacher, peer-to-peer interactions can have significant advantages over tutor-to-student interactions. Robovie was the first fully autonomous robot to be introduced into an elementary school (*70*). It was an English-speaking robot targeting two grades (first and sixth) of Japanese children. Through field trials conducted over 2 weeks, improvements in English language skills were observed in some children. In one case, longer periods of attention on learning tasks, faster responses, and more accurate responses were shown with a peer robot compared with an identical-looking tutor robot (*19*). A long-term primary school study showed that a peer-like humanoid robot able to personalize the interaction could increase child learning of novel subjects (*48*). Often, the robot is presented as a more knowledgeable peer, guiding the student along a learning trajectory that is neither too easy nor too challenging. However, the role of those robots sometimes becomes ambiguous (tutor versus peer), and it is difficult to place one above the other in general. Learning companions (*71*), which offer motivational support but otherwise are not tutoring, are also successful cases of a peer-like robot.

Robot as novice

Considerable educational benefits can also be obtained from a robot that takes the role of a novice, allowing the student to take on the role of an instructor that typically improves confidence while, at the same time, establishing learning outcomes. This is an instance of learning by teaching, which is widely known in human education, also referred to as the protégé effect (*72*). This process involves the learner making an effort to teach the robot, which has a direct impact on their own learning outcomes.

The care-receiving robot (CRR) was the first robot designed with the concept of a teachable robot for education (*73*). A small humanoid robot introduced into English classes improved the vocabulary learning of 3- to 6-year-old Japanese children (*5*). The robot was designed to make deliberate errors in English vocabulary but could be corrected through instruction by the children. In addition, CRR was shown to engage children more than alternative technology, which eventually led to the release of a commercial product based on the principle of a robot as a novice (*74*).

This novice role can also be used to teach motor skills. The CoWriter project explored the use of a teachable robot to help children improve their handwriting skills (*13*). A small humanoid robot in conjunction with a touch tablet helped children who struggled with handwriting to improve their fine motor skills. Here, the children taught the robot, who initially had very poor handwriting, and in the process of doing so, the children reflected on their own writing and showed improved motor skills (*13*). This suggests that presenting robots as novices has potential to develop meta-cognitive skills in learners, because the learners are committing to instructing the learning material, requiring a higher level of understanding of the material and an understanding of the internal representations of their robot partner.

In our meta-analysis, the robot was predominantly used as a tutor (48%), followed by a role as teacher (38%). In only 9% of studies was the robot presented as a peer or novice (Fig. 1B). The robot was often used to offer one-to-one interactions (65%), with the robot used in a one-to-many teaching scenario in only 30% of the studies (Fig. 1C). In 5%, the robot had mixed interactions, whereby, for example, it first taught more than one student and then had one-on-one interactions during a quiz.

DISCUSSION

Although an increasing number of studies confirm the promise of social robots for education and tutoring, this Review also lays bare a number of challenges for the field. Robots for learning, and social robotics in general, require a tightly integrated endeavor. Introducing these technologies into educational practice involves solving technical challenges and changing educational practice.

With regard to the technical challenges, building a fluent and contingent interaction between social robots and learners requires the seamless integration of a range of processes in artificial intelligence and robotics. Starting with the input to the system, the robot needs a sufficiently correct interpretation of the social environment

for it to respond appropriately. This requires significant progress in constituent technical fields, such as speech recognition and visual social signal processing, before the robot can access the social environment. Speech recognition, for example, is still insufficiently robust to allow the robot to understand spoken utterances from young children. Although these shortcomings can be resolved by using alternative input media, such as touch screens, this does place a considerable constraint on the natural flow of the interaction. For robots to be autonomous, they must make decisions about which actions to take to scaffold learning. Action selection is a challenging domain at best and becomes more difficult when dealing with a pedagogical environment, because the robot must have an understanding of the learner's ability and progress to allow it to choose appropriate actions. Finally, the generation of verbal and nonverbal output remains a challenge, with the orchestrated timing of verbal and nonverbal actions a prime example. In summary, social interaction requires the seamless functioning of a wide range of cognitive mechanisms. Building artificial social interaction requires the artificial equivalent of these cognitive mechanisms and their interfaces, which is why artificial social interaction is perhaps one of the most formidable challenges in artificial intelligence and robotics.

Introducing social robots in the school curriculum also poses a logistical challenge. The generation of content for social robots for learning is nontrivial, requiring tailor-made material that is likely to be resource-intensive to produce. Currently, the value of the robot lies in tutoring very specific skills, such as mathematics or handwriting, and it is unlikely that robots can take up the wide range of roles a teacher has, such as pedagogical and carer roles. For the time being, robots are mainly deployed in elementary school settings. Although some studies have shown the efficacy of tutoring adolescents and adults, it is unclear whether the approaches that work well for younger children transfer to tutoring older learners.

Introducing robots might also carry risks. For example, studies of ITS have shown that children often do not make the best use of on-demand support and either rely too much on the help function or avoid using help altogether, both resulting in suboptimal learning. Although strategies have been explored to mitigate this particular problem in robots (*4*), there might be other problems specific to social robots that still need to be identified and for which solutions will be needed.

Social robots have, in the broadest sense, the potential to become part of the educational infrastructure, just as paper, white boards, and computer tablets have. Next to the functional dimension, robots also offer unique personal and social dimensions. A social robot has the potential to deliver a learning experience tailored to the learner, supporting and challenging students in ways unavailable in current resource-limited educational environments. Robots can free up precious time for human teachers, allowing the teacher to focus on what people still do best: providing a comprehensive, empathic, and rewarding educational experience.

Next to the practical considerations of introducing robots in education, there are also ethical issues. How far do we want the education of our children to be delegated to machines, and social robots in particular? Overall, learners are positive about their experience with robots for learning, but parents and teaching staff adopt a more cautious attitude (*75*). There is much to gain from using robots, but what do we stand to lose? Might robots lead to an impoverished learning experience where what is technologically possible is prioritized over what is actually needed by the learner?

Notwithstanding, robots show great promise when teaching restricted topics, with effect sizes on cognitive outcomes almost matching those of human tutoring. This is remarkable, because our meta-analysis gathered results from a wide range of countries using different robot types, teaching approaches, and deployment contexts. Although the use of robots in educational settings is limited by technical and logistical challenges for now, the benefits of physical embodiment may lift robots above competing learning technologies, and classrooms of the future will likely feature robots that assist a human teacher.

REFERENCES AND NOTES

1. N. C. Krämer, G. Bente, Personalizing e-Learning. The social effects of pedagogical agents. *Educ. Psychol. Rev.* **22**, 71–87 (2010).
2. J. A. Kulik, J. D. Fletcher, Effectiveness of intelligent tutoring systems: A meta-analytic review. *Rev. Educ. Res.* **86**, 42–78 (2016).
3. J. Kennedy, P. Baxter, E. Senft, T. Belpaeme, in *Proceedings of the International Conference on Social Robotics* (Springer, 2015), pp. 327–336.
4. A. Ramachandran, A. Litoiu, B. Scassellati, in *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction* (IEEE, 2016), pp. 247–254.
5. F. Tanaka, S. Matsuzoe, Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *J. Hum. Robot Interact.* **1**, 78–95 (2012).
6. I. Leite, C. Martinho, A. Paiva, Social robots for long-term interaction: A survey. *Int. J. Soc. Robot.* **5**, 291–308 (2013).
7. J. Han, *Robot-Aided Learning and r-Learning Services* (INTECH Open Access Publisher, 2010).
8. O. Mubin, C. J. Stevens, S. Shahid, A. Al Mahmud, J.-J. Dong, A review of the applicability of robots in education. *J. Technol. Educ. Learning* **1**, 1–7 (2013).
9. J. Gorham, The relationship between verbal teacher immediacy behaviors and student learning. *Commun. Educ.* **37**, 40–53 (1988).
10. P. L. Witt, L. R. Wheeless, M. Allen, A meta-analytical review of the relationship between teacher immediacy and student learning. *Commun. Monogr.* **71**, 184–207 (2004).
11. M. Saerbeck, T. Schut, C. Bartneck, M. D. Janse, 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*, CHI'10 (ACM, 2010), pp. 1613–1622.
12. V. Girotto, C. Lozano, K. Muldner, W. Burleson, E. Walker, Lessons learned from in-school use of rtag: A robo-tangible learning environment, in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (ACM, 2016), pp. 919–930.
13. D. Hood, S. Lemaignan, P. Dillenbourg, When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting, in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 83–90.
14. A. Litoiu, B. Scassellati, Robotic coaching of complex physical skills, in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 211–212.
15. J. Fasola, M. Mataric, A socially assistive robot exercise coach for the elderly. *J. Hum. Robot Interact.* **2**, 3–32 (2013).
16. A. Kulkarni, A. Wang, L. Urbina, A. Steinfeld, B. Dias, in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, 2016), pp. 461–462.
17. B. Scassellati, J. Brawer, K. Tsui, S. N. Gilani, M. Malzkuhn, B. Manini, A. Stone, G. Kartheiser, A. Merla, A. Shapiro, D. Traum, L. Petitto, Teaching language to deaf infants with a robot and a virtual human, in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 21 to 26 April 2018, Montréal, Canada (ACM, 2018).
18. R. A. Richert, M. B. Robb, E. I. Smith, Media as social partners: The social nature of young children's learning from screen media. *Child Dev.* **82**, 82–95 (2011).
19. C. Zaga, M. Lohse, K. P. Truong, V. Evers, The effect of a robot's social character on children's task engagement: Peer versus tutor, in *International Conference on Social Robotics* (Springer, 2015), pp. 704–713.
20. J. Kennedy, P. Baxter, T. Belpaeme, Comparing robot embodiments in a guided discovery learning interaction with children. *Int. J. Soc. Robot.* **7**, 293–308 (2015).
21. C. D. Kidd, C. Breazeal, Effect of a robot on user perceptions, in *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004 (IROS 2004)* (IEEE, 2004), vol. 4, pp. 3559–3564.
22. J. Wainer, D. J. Feil-Seifer, D. A. Shell, M. J. Mataric, in *Proceedings of the 16th IEEE International Symposium on Robot and Human interactive Communication, RO-MAN* (IEEE, 2007), pp. 872–877.
23. H. Köse, P. Uluer, N. Akalın, R. Yorgancı, A. Özkul, G. Ince, The effect of embodiment in sign language tutoring with assistive humanoid robots. *Int. J. Soc. Robot.* **7**, 537–548 (2015).

24. A. Powers, S. Kiesler, S. Fussell, C. Torrey, Comparing a computer agent with a humanoid robot, in *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction* (IEEE, 2007), pp. 145–152.
25. J. Li, The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *Int. J. Hum. Comput. Stud.* **77**, 23–37 (2015).
26. W. A. Bainbridge, J. W. Hart, E. S. Kim, B. Scassellati, The benefits of interactions with physically present robots over video-displayed agents. *Int. J. Soc. Robot.* **3**, 41–52 (2011).
27. D. Leyzberg, S. Spaulding, M. Toneva, B. Scassellati, The physical presence of a robot tutor increases cognitive learning gains, in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, CogSci 2012 (2012), pp. 1882–1887.
28. C. D. Kidd, C. Breazeal, A robotic weight loss coach, in *Proceedings of the National Conference on Artificial Intelligence* (MIT Press, 2007), vol. 22, pp. 1985–1986.
29. J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, T. Belpaeme, Child speech recognition in human-robot interaction: Evaluations and recommendations, in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (ACM/IEEE, 2017), pp. 82–90.
30. P. Baxter, R. Wood, T. Belpaeme, A touchscreen-based 'sandtray' to facilitate, mediate and contextualise human-robot social interaction, in *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2012), pp. 105–106.
31. A. Ramachandran, C.-M. Huang, B. Scassellati, Give me a break! Personalized timing strategies to promote learning in robot-child tutoring, in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2017), pp. 146–155.
32. D. Szafir, B. Mutlu, Pay attention! Designing adaptive agents that monitor and improve user engagement, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'12 (ACM, 2012), pp. 11–20.
33. I. Leite, M. McCoy, D. Ullman, N. Salomons, B. Scassellati, Comparing models of disengagement in individual and group interactions, in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 99–105.
34. A. Coninx, P. Baxter, E. Oleari, S. Bellini, B. Bierman, O. B. Henkemans, L. Cañamero, P. Cosi, V. Enescu, R. Ros Espinoza, A. Hiolle, R. Humbert, B. Kiefer, Towards long-term social child-robot interaction: Using multi-activity switching to engage young users. *J. Hum. Robot Interact.* **5**, 32–67 (2016).
35. S. Lemaignan, F. Garcia, A. Jacq, P. Dillenbourg, From real-time attention assessment to "with-me-ness" in human-robot interaction, in *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction* (IEEE, 2017).
36. I. Leite, G. Castellano, A. Pereira, C. Martinho, A. Paiva, Empathic robots for long-term interaction. *Int. J. Soc. Robot.* **6**, 329–341 (2014).
37. A. Ramachandran, C.-M. Huang, E. Gartland, B. Scassellati, Thinking aloud with a tutoring robot to enhance learning, in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2018), pp. 59–68.
38. C.-M. Huang, B. Mutlu, Modeling and evaluating narrative gestures for humanlike robots, in *Proceedings of the Robotics: Science and Systems Conference*, RSS'13 (2013).
39. C.-M. Huang, B. Mutlu, The repertoire of robot behavior: Enabling robots to achieve interaction goals through social behavior. *J. Hum. Robot Interact.* **2**, 80–102 (2013).
40. J. Kennedy, P. Baxter, T. Belpaeme, The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning, in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2015), pp. 67–74.
41. E. Yadollahi, W. Johal, A. Paiva, P. Dillenbourg, When deictic gestures in a robot can harm child-robot collaboration, in *Proceedings of the 17th ACM Conference on Interaction Design and Children* (ACM, 2018), pp. 195–206.
42. G. Gordon, C. Breazeal, Bayesian active learning-based robot tutor for children's word-reading skills, in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, AAAI-15 (2015).
43. D. Leyzberg, S. Spaulding, B. Scassellati, Personalizing robot tutors to individual learning differences, in *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2014).
44. T. Schodde, K. Bergmann, S. Kopp, Adaptive robot language tutoring based on Bayesian knowledge tracing and predictive decision-making, in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2017), pp. 128–136.
45. J. Janssen, C. van der Wal, M. Neerincx, R. Looije, Motivating children to learn arithmetic with an adaptive robot game, in *Proceedings of the Third international conference on Social Robotics* (ACM, 2011), pp. 153–162.
46. O. A. Blanson Henkemans, B. P. Bierman, J. Janssen, M. A. Neerincx, R. Looije, H. van der Bosch, J. A. van der Giessen, Using a robot to personalise health education for children with diabetes type 1: A pilot study. *Patient Educ. Couns.* **92**, 174–181 (2013).
47. G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, C. Breazeal, Affective personalization of a social robot tutor for children's second language skills, in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (AAAI, 2016), pp. 3951–3957.
48. P. Baxter, E. Ashurst, R. Read, J. Kennedy, T. Belpaeme, Robot education peers in a situated primary school study: Personalisation promotes child learning. *PLOS ONE* **12**, e0178126 (2017).
49. D. Leyzberg, E. Avrunin, J. Liu, B. Scassellati, Robots that express emotion elicit better human teaching, in *Proceedings of the 6th International Conference on Human-Robot Interaction* (ACM, 2011), pp. 347–354.
50. C. D. Kidd, "Designing for long-term human-robot interaction and application to weight loss," thesis, Massachusetts Institute of Technology (2008).
51. The meta-analysis data are available at https://tinyurl.com/ybuyz5vn.
52. B. Bloom, M. Engelhart, E. Furst, W. Hill, D. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain* (Donald McKay, 1956).
53. D. Krathwohl, B. Bloom, B. Masia, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook II: The Affective Domain* (Donald McKay, 1964).
54. D. R. Krathwohl, A revision of bloom's Taxonomy: An overview. *Theory Pract.* **41**, 212–218 (2002).
55. https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php
56. M. W. Lipsey, D. B. Wilson, *Practical Meta-Analysis* (Sage Publications, Inc, 2001).
57. C.-M. Huang, B. Mutlu, Learning-based modeling of multimodal behaviors for humanlike robots, in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2014), pp. 57–64.
58. B. S. Bloom, The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educ. Res.* **13**, 4–16 (1984).
59. K. VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* **46**, 197–221 (2011).
60. T. W. Draper, W. W. Clayton, Using a personal robot to teach young children. *J. Genet. Psychol.* **153**, 269–273 (1992).
61. M. Imai, T. Ono, H. Ishiguro, Physical relation and expression: Joint attention for human-robot interaction. *IEEE Trans. Ind. Electron.* **50**, 636–643 (2003).
62. B. Mutlu, J. Forlizzi, J. Hodgins, A storytelling robot: Modeling and evaluation of human-like gaze behavior, in *Humanoid Robots, 2006 6th IEEE-RAS International Conference* (IEEE, 2006), pp. 518–523.
63. Z.-J. You, C.-Y. Shen, C.-W. Chang, B.-J. Liu, G.-D. Chen, A robot as a teaching assistant in an English class, in *Proceedings of the Sixth International Conference on Advanced Learning Technologies* (IEEE, 2006), pp. 87–91.
64. T. Belpaeme, P. Vogt, R. Van den Berghe, K. Bergmann, T. Göksun, M. De Haas, J. Kanero, J. Kennedy, A. C. Küntay, O. Oudgenoeg-Paz, F. Papadopoulos, Guidelines for designing social robots as second language tutors. *Int. J. Soc. Robot.* **10**, 1–17 (2018).
65. J.-H. Han, M.-H. Jo, V. Jones, J.-H. Jo, Comparative study on the educational use of home robots for children. *J. Inform. Proc. Syst.* **4**, 159–168 (2008).
66. J. Movellan, F. Tanaka, I. Fasel, C. Taylor, P. Ruvolo, M. Eckhardt, The RUBI project: A progress report, in *Proceedings of the Second ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2007).
67. F. Tanaka, A. Cicourel, J. R. Movellan, Socialization between toddlers and robots at an early childhood education center. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17954–17958 (2007).
68. J. R. Movellan, M. Eckhardt, M. Virnes, A. Rodriguez, Sociable robot improves toddler vocabulary skills, in *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2009), pp. 307–308.
69. M. Alemi, A. Meghdari, M. Ghazisaedy, Employing humanoid robots for teaching English language in Iranian junior high-schools. *Int. J. Humanoid Robot.* **11**, 1450022 (2014).
70. T. Kanda, T. Hirano, D. Eaton, H. Ishiguro, Interactive robots as social partners and peer tutors for children: A field trial. *Hum. Comput. Interact.* **19**, 61–64 (2004).
71. N. Lubold, E. Walker, H. Pon-Barry, Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion, in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction* (IEEE Press, 2017), pp. 255–262.
72. C. C. Chase, D. B. Chin, M. A. Oppezzo, D. L. Schwartz, Teachable agents and the protégé effect: Increasing the effort towards learning. *J. Sci. Educ. Technol.* **18**, 334–352 (2009).
73. F. Tanaka, T. Kimura, The use of robots in early education: A scenario based on ethical consideration, in *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication* (IEEE, 2009), pp. 558–560.
74. F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, K. Hayashi, Pepper learns together with children: Development of an educational application, in *IEEE-RAS 15th International Conference on Humanoid Robots*, HUMANOIDS 2015 (IEEE, 2015), pp. 270–275.
75. J. Kennedy, S. Lemaignan, T. Belpaeme, The cautious attitude of teachers towards social robots in schools, in *Proceedings of the Robots 4 Learning Workshop at RO-MAN 2016* (2016).
76. I. Leite, A. Pereira, G. Castellano, S. Mascarenhas, C. Martinho, A. Paiva, Social robots in learning environments: A case study of an empathic chess companion, in *Proceedings of the International Workshop on Personalization Approaches in Learning Environments* (2011).
77. R. R. Hake, Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.* **66**, 64–74 (1998).
78. C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* **1**, 71–81 (2009).

79. R. F. Ferguson, *The Tripod Project Framework* (Tripod, 2008).
80. M. Alemi, A. Meghdari, M. Ghazisaedy, The impact of social robotics on l2 learners' anxiety and attitude in English vocabulary acquisition. *Int. J. Soc. Robot.* **7**, 523–535 (2015).
81. M. Fridin, Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Comput. Educ.* **70**, 53–64 (2014).
82. D. Jolliffe, D. P. Farrington, Development and validation of the basic empathy scale. *J. Adolesc.* **29**, 589–611 (2006).

Acknowledgments: We are grateful to E. Ashurst for support in collecting the data for the meta-analysis. **Funding:** This work is partially funded by the H2020 L2TOR project (688014), Japan Society for the Promotion of Science KAKENHI (15H01708), and NSF award 1139078.

Author contributions: All authors contributed equally to the manuscript; T.B. and J.K. contributed to the meta-analysis. **Competing interests:** J.K. is a research scientist at Disney Research. **Data and materials availability:** The meta-analysis data are available at https://tinyurl.com/ybuyz5vn.

Submitted 31 March 2018
Accepted 23 July 2018
Published 15 August 2018
10.1126/scirobotics.aat5954

Citation: T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka, Social robots for education: A review. *Sci. Robot.* **3**, eaat5954 (2018).

Social robots for education: A review

Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati and Fumihide Tanaka

Sci. Robotics **3**, eaat5954.
DOI: 10.1126/scirobotics.aat5954

Using a Robot Peer to Encourage the Production of Spatial Concepts in a Second Language

Conference Paper · December 2018

DOI: 10.1145/3284432.3284433

7 authors, including:

Christopher David Wallbridge
University of Plymouth
5 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Rianne van den Berghe
Utrecht University
7 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Daniel Hernandez Garcia
University of Plymouth
14 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Séverin Lemaignan
Bristol Robotics Laboratory
76 PUBLICATIONS **1,166** CITATIONS
SEE PROFILE

Using a Robot Peer to Encourage the Production of Spatial Concepts in a Second Language

Christopher D. Wallbridge
University of Plymouth
Plymouth, UK
christopher.wallbridge@plymouth.ac.uk

Rianne van den Berghe
Utrecht University
Utrecht, Netherlands
m.a.j.vandenberghe@uu.nl

Daniel Hernández García
University of Plymouth
Plymouth, UK
daniel.hernandez@plymouth.ac.uk

Junko Kanero
Koç University
Istanbul, Turkey
jkanero@ku.edu.tr

Séverin Lemaignan
Bristol Robotics Laboratory
Bristol, UK
severin.lemaignan@brl.ac.uk

Charlotte Edmunds
University of Plymouth
Plymouth, UK
charlotte.edmunds@plymouth.ac.uk

Tony Belpaeme
Ghent University/University of Plymouth
Ghent, Belgium
tony.belpaeme@ugent.be

ABSTRACT

We conducted a study with 25 children to investigate the effectiveness of a robot measuring and encouraging production of spatial concepts in a second language compared to a human experimenter. Productive vocabulary is often not measured in second language learning, due to the difficulty of both learning and assessing productive learning gains. We hypothesized that a robot peer may help assessing productive vocabulary. Previous studies on foreign language learning have found that robots can help to reduce language anxiety, leading to improved results. In our study we found that a robot is able to reach a similar performance to the experimenter in getting children to produce, despite the person's advantages in social ability, and discuss the extent to which a robot may be suitable for this task.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Social and professional topics** → **Assistive technologies**; • **Computing methodologies** → Natural language processing; Cognitive robotics;

KEYWORDS

Robot Assisted Language Learning; Assessment; Second Language Learning

HAI '18, December 15–18, 2018, Southampton, United Kingdom

ACM ISBN 978-1-4503-5953-5/18/12…$15.00
https://doi.org/10.1145/3284432.3284433

ACM Reference Format:
Christopher D. Wallbridge, Rianne van den Berghe, Daniel Hernández García, Junko Kanero, Séverin Lemaignan, Charlotte Edmunds, and Tony Belpaeme. 2018. Using a Robot Peer to Encourage the Production of Spatial Concepts in a Second Language. In *6th International Conference on Human-Agent Interaction (HAI '18), December 15–18, 2018, Southampton, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3284432.3284433

1 INTRODUCTION

Learning the language of a new home region is vital for migrant children. It is beneficial for them to integrate with their peers, and necessary to prevent them from falling behind in school. Children need the opportunity to practice their language skills, but it may be difficult if no one at home is able to speak the language of the host region. Finding qualified teachers or tutors that know both the new language and the language of children's old homeland can also be challenging. With robots we may be able to support children's language learning needs.

When learning a second language (L2), it is difficult to master vocabulary both receptively and productively. L2 learners may find themselves capable of understanding the L2, while still struggling to produce L2 words. Indeed, previous research has shown that receptive vocabulary tends to be bigger than productive vocabulary in first language (L1) [8, 11], and that L2 learners obtain lower scores on productive tests as compared to receptive tests [14]. Thus, people are able to recognize more words than they can produce, both in their L1 and L2. This has been formalised into a hierarchy for word knowledge by Laufer et al. [9], based on knowing the words passively or actively and in being able to recognize them or recall them. The hierarchy is as follows, from easiest to most difficult: passive recognition → active recognition → passive recall → active recall. These are defined as follows:

- *Passive recognition* - The student is able to select the L1 word from a choice of words when provided the word in L2.

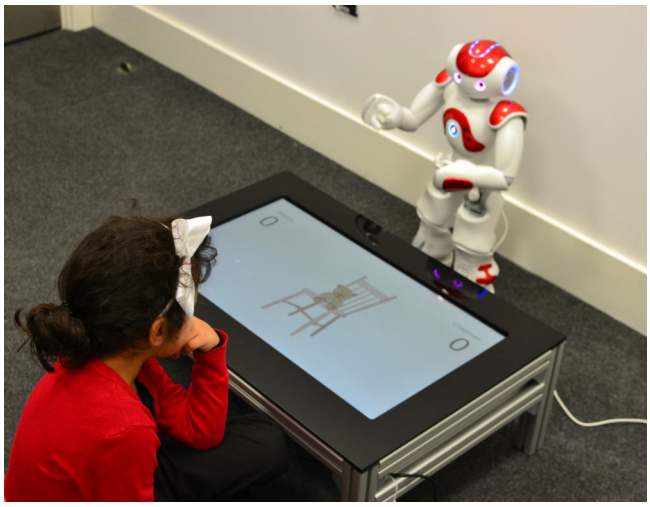


Figure 1: A child interacting with the robot in our study. The agent – in this case a robot – stands opposite from the child. An interactive table displays an image of a teddy bear and a chair. The child must use a word from a second language to describe the position of the bear in relation to the chair.

- *Active recognition* - The student is able to select the L2 word from a choice of words when provided the word in L1.
- *Passive recall* - The student is able to give the meaning of a word in L1 when provided the word in L2.
- *Active recall* - The student is able to give the L2 word when provided the word in L1.

This poses a challenge for L2 vocabulary interventions in which the trainer wants to assess the trainee's learning gains: L2 learners have difficulty learning the words productively (i.e. learning to produce foreign words), and will struggle to actively recall newly learned L2 words. There are several tests to assess an L2 learner's productive vocabulary, including assessments in which the participant has to describe pictures (e.g., the Expressive Vocabulary Test [18], the Expressive One-Word Picture Vocabulary Test [5], or the Clinical Evaluation of Language Fundamentals Test [17]), writing tests in which the learner has to fill in the blank (e.g., the Productive Vocabulary Levels Test [10]), or, for very young children, parental or teacher reports [4].

In many situations, it may not be possible to use one of these tests. For example, when the words learned concern abstract concepts, which cannot be easily depicted, it is not possible to use a picture test. If the learner is illiterate, one cannot use a writing test. Parents or teachers may struggle to report the child's L2 if they do not speak that language themselves. To further complicate the issue, producing L2 words may be intimidating for L2 learners. Even if the learner is able to produce the word, they may not produce it due to anxiety of pronouncing the word incorrectly [13].

A social robot may help overcome some of the issues described above in assessing L2 learner's vocabulary. While not being able to solve by itself the issue of vocabulary being more difficult to learn productively than receptively, a social robot may help in innovating novel ways to assess L2 vocabulary, or in reducing L2 anxiety in L2 vocabulary test settings. A robot may be less intimidating than an adult assessor, especially for young children, encouraging more speech production. This study evaluates whether school children may produce more L2 words in a productive L2 vocabulary test when playing with a social robot than with an adult. Below, we discuss relevant robot-assisted language learning (RALL) studies before detailing our study.

2 PREVIOUS WORK

RALL has been found to be effective in reducing foreign language anxiety (FLA), and teaching robots are able to improve oral skills of young students learning English as a foreign language [1]. Alemi et al. [2] performed a study using a robot teaching assistant. In the study, Persian-speaking students in Iran were taught English. A survey of the students showed that those who learned from the robot were significantly less anxious compared to the control group that did not have the robot. While a number of factors were thought to contribute to this reduction in anxiety, the authors claimed a major reason to be intentional mistakes the robot made. The mistakes not only gave the students a chance to correct the robot, but also made them less afraid of making errors of their own.

When looking at speaking skills, the focus can not just be on vocabulary gains, but pronunciation as well. Lee et al. [12] conducted a series of lessons to help Korean children from grades 3 to 5 (roughly 8 to 10 years old) learn English. In South Korea children start learning English from grade 3. As part of a lesson series they were given a pronunciation training with a robot, that used a lexicon that included often confused phonemes, so that the robot could correct the child's pronunciation. It was reported that the children's speaking skills improved significantly with a large effect size when measured by a teacher. As well as the improvement in speaking skills all three affective factors – interest, confidence and motivation – all improved significantly.

Instances of robots acting as care-receivers also occur in RALL. In a study by Tanaka and Matsuzoe [16], Japanese children were given the role of teaching English verbs to a NAO robot. The children had to guide the robot's arm to act out the target verbs, e.g. brushing teeth. In a comprehension post-test the children answered correctly more often with words they had taught the robot than those learnt during a regular verb-learning game. While the robot only learned from 'Direct' teaching, where the child was guiding the motion of the robot, there was a high frequency of verbal teaching using English.

We can see that there are many instances where RALL is able to assist in teaching an L2 to students. Many of these show a reduction in FLA and increase in confidence and willingness to learn in the students. In all these cases, however, they use the robot to teach, whether directly in the role of teacher or acting as a care receiver or assistant. Robots were not used in assessment, and in most cases the tests performed were aimed at measuring the comprehension of the L2 words that were being taught. We want to explore the possibility of using a robot to assess the L2 production of children. Due to the reported reductions in anxiety and increase in confidence when using a robot, we may see an increase in the amount of production.

3 STUDY DESIGN

This study was conducted at a local school with English-speaking 5- to 6-year-old children. We decided to teach spatial language, more specifically spatial prepositions, because while those concepts are more abstract than physical objects, we can still represent them using images. Spatial language itself is also particularly challenging to L2 learners as the meaning can often differ depending on context and the referent. Every morning, five children were randomly selected to participate in the study for that day and assigned a condition, balanced across gender. These five children were first given a French lesson before playing our production quiz game on an interactive table [3] individually throughout the rest of the day (Figure 1). An agent (robot or experimenter depending on our condition) is placed opposite to the child and gives instructions and encouragement to the children. The interactive table displays an image of a teddy bear and a chair. The child would have to use one of the French words taught to describe the position of the bear relative to the chair.

As well as the teacher three experimenters were involved in the study:

(1) *Lead Experimenter* - The lead experimenter acted as the interaction point for the children outside of the one to one sessions. Either the lead experimenter or the wizard was required to be in the presence of the child while outside their classroom. The lead experimenter was certified in the children's health and well being, and was there to ensure the health and safety of the children as required by the school.
(2) *Wizard Experimenter* - The wizard experimenter controlled the robot remotely via a laptop interface. The wizard experimenter was also certified in the children's health and well being, but had minimal interaction with the children so as to minimise interference during the study.
(3) *Blind Experimenter* - The blind experimenter facilitated the interactions before the main study began, provided the comprehension test and acted as the agent in the child-human condition. The blind experimenter was unaware of the purpose of the study to reduce influencing the outcome.

3.1 Hypothesis

With our study we wanted to test the following hypothesis:

H The presence of a robot will allow children to produce more spatial words verbally in an L2 than when working with a human experimenter.

3.2 Teaching

The children were taught five French words: *Nounours* (Teddy Bear), *chaise* (chair), *devant* (in front of), *sur* (on), *sous* (under). Of these, the first two were supporting words and the last three were the target words for the study. The content of the lesson was created and taught by a professional French teacher, with a goal of enabling the children to produce these words after one lesson. We decided to use a professional teacher as we did not want a robot teacher that would also influence our results. It has also been shown that human teachers can still outperform a robot teacher [7]. The lead experimenter acted as a teacher's assistant. The children were taught in groups of five. The lesson was designed to last 30 minutes.

The teacher started the lesson by introducing the children to the support words. At all stages the children were encouraged to repeat any French words they heard. The children were taught a song that used the three target words and hand gestures to go along with them. After singing, the children would position themselves relative to the chair based on the words announced by the teacher. The children were then each given a teddy bear and repeated the process with the bear. The children then played a game of 'Telephone'. In this game one child was first given one of the target words, and each child would whisper the word to the next child down the line until the last child. The last child would announce to the rest of the group the word they heard. The game was repeated several times with the children re-organised into a different order so that the announcing child changed each time. This was followed by a game of 'Corners'. In each corner of the lesson area, a teddy was placed in a position relative to a chair that referred to one of the target words. The children were then encouraged to sing and move around until the teacher would stop them, and say one of the target words. The children then had to move to the relevant corner and say the word three times. Variants of this game were then played in teams with the chairs lined up, and then individually. Finally each child was told to say one of the target words and then go stand by the correct chair. The lesson wrapped up with one more repetition of the song they had been taught near the beginning.

During the interaction we also established any prior knowledge in the target language. They were split into the following categories:

(1) *No Exposure* - The children have not been exposed to any French, other than potentially those used in popular culture e.g. *C'est la vie*.
(2) *Beginner* - The child has potentially received some lessons in French and knows simple phrases that do not include our target words e.g. *Je m'appelle John*.
(3) *Intermediate* - The child has knowledge of French, including our target words.
(4) *Advanced* - The child has an intricate knowledge of French, and is able to produce words with a high capability or are fluent.

Children of intermediate or advanced knowledge were excluded from the data analysis. 25 children took part in our study of which three were excluded from the analysis of results, leaving 22 children.

3.3 Individual Interactions

Upon completing another familiarity task and a 10 minute activity with the robot–that required the child to describe the position of objects to the robot in English–a comprehension test was administered by a blind experimenter who was unaware of the purpose of the study (Figure 2). This served as a small refresher of what the children had learned earlier in the day, as well as allows us to establish a baseline for the efficacy of the lesson. For the comprehension test there were 6 sheets with 3 images each (representing the 3 target words), placed on the left, in the centre or on the right. Together, the 6 sheets covered all possible permutations of the 3 target words (*devant*, *sur*, *sous*) with each of the 3 positions. The images were similar but not the same as the ones used for the production quiz questions. For each sheet the experimenter asked the child to point at the picture that matches the statement (see below). If the

Figure 2: A child being administered the comprehension test before moving onto the main production quiz.

Figure 3: The 'wizard' experimenter was positioned behind the child to minimise interaction between them.

child pointed to the wrong picture they were allowed to try again until they pointed to the correct image. We repeated each target word twice to account for guessing and to ensure they weren't just picking based on location on the question sheet. The statements and their order were the same for every child:

(1) Le nounours est sous la chaise.
(2) Le nounours est devant la chaise.
(3) Le nounours est sur la chaise.
(4) Le nounours est devant la chaise.
(5) Le nounours est sur la chaise.
(6) Le nounours est sous la chaise.

The child then played the production quiz with either the robot or the blind experimenter based on the group they were in (child-robot or child-human). In both conditions, the production quiz was displayed on the sandtray. The robot was controlled through a Wizard-of-Oz interface, with the 'wizard' sat behind the child, out of sight, so as to minimise effects on the child (Figure 3). The rules of the game were explained by the agent (blind experimenter or robot). The child was sat in front of the sandtray upon which the production quiz game was displayed. The agent sat opposite the child. The sandtray displayed an image of the teddy bear in a position relative to the chair, and the agent or child must answer "Où est le nounours?" (Where is the teddy bear?). The agent was to give the answer in the form "sur/sous/devant la chaise", but any answer given by the child that included one of the target words 'sur', 'sous' or 'devant' was accepted. Each correct answer scored a point. If either the question was answered correctly or both the child and the agent answered incorrectly then the production quiz moved onto the next question. If the child did not answer after a short period then the agent would give encouragement in proceeding levels:

(1) Encourage the child to guess e.g. "Just have a guess".
(2) Targeted encouragement, such as asking them to remember the lesson from the morning.
(3) The agent will attempt the question.
    - If the child was ahead on points then the agent (adult/robot) would answer correctly so as to keep up an appearance of a challenging opponent in the game.
    - If the child was level or behind the agent (adult/robot) then the agent would answer incorrectly to demonstrate a willingness to answer even if wrong.

If the child still did not have a guess after all stages then the game proceeded as if they had answered incorrectly. The agent began the production quiz after explaining how to play by answering the first question correctly. There were nine subsequent questions which we expected the child to answer, three for each target word.

4 RESULTS

4.1 Participants

25 children took part in our study of which three were excluded from our analysis of results leaving us with 22 children. 11 Children were in the Human Condition (4 Female) and 11 in the Robot Condition (6 Female). There were 11 5 year olds (6 Female) and 11 6 year olds (4 Female). Of these children two had an L1 other than English (1 Female), but their English level was high enough to still participate.

4.2 Comprehension

We scored the comprehension test by taking the maximum attempts per question (3) and subtracting the number of attempts they took to get the correct answer. This meant each question was scored between 0 and 2, giving a maximum possible score of 12 on the comprehension test. The mean score for the comprehension test was 8.5 (SD=1.92). In the Human condition the children averaged 8.27 (SD=2.20) at the comprehension test while in the Robot condition the children averaged 8.72 (SD=1.68). Using a Welch Two Sample t-test, no significant difference between the two conditions was found ($t = 0.55$, $df = 18.72$ $p = 0.59$). This shows that the groups between our two conditions were roughly equal in ability before beginning the

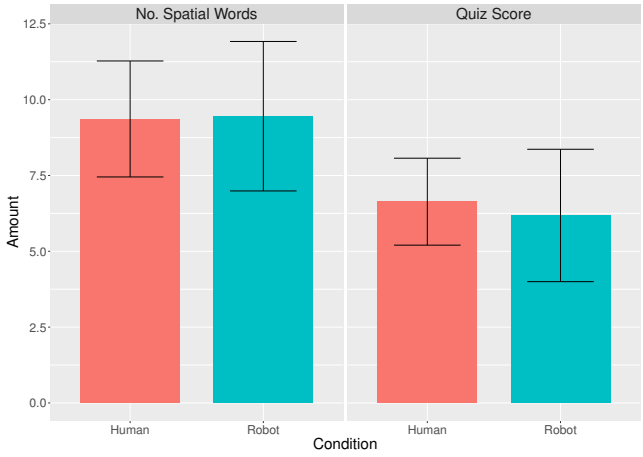


Figure 4: Analysis of L2 spatial words used during the production quiz. Left: spatial words used without additional prompting to attempt the question; right: number of correct words said by the children during the production quiz. In both cases no significant difference was found between the robot and adult conditions. Error bars are showing the standard deviation.

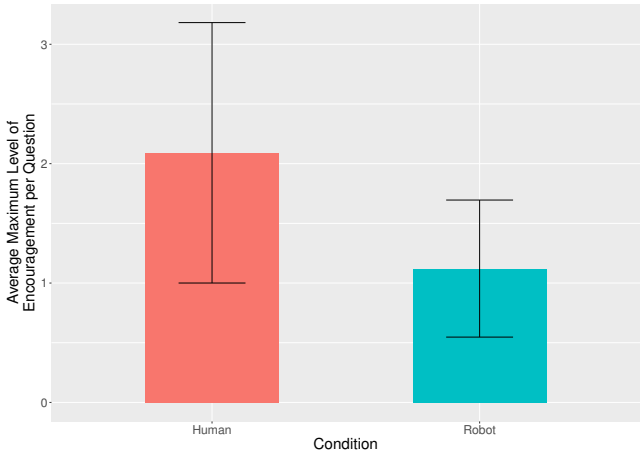


Figure 5: Analysis between participants of the average maximum level of encouragement reached across conditions. A significant difference is seen between the two conditions, Human and Robot. Error bars are showing the standard deviation.

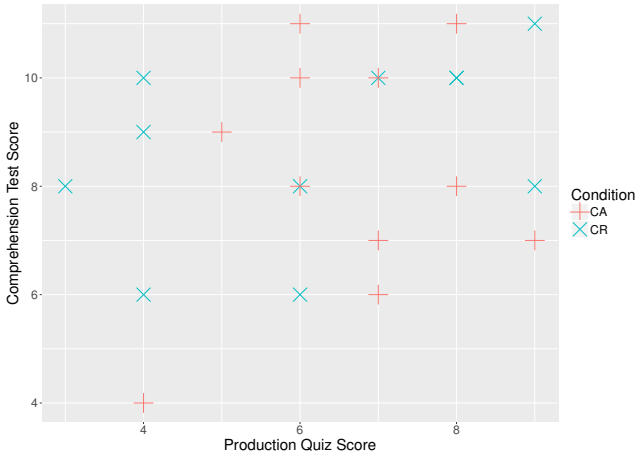


Figure 6: A comparison between the score in the production quiz and the score on the comprehension test. No significant correlation was found.

production quiz. The scores remained consistent throughout the test, with no learning effect seen when the first half and the second half of the comprehension test were compared (first half: mean=4.5, SD=1.26; second half: mean=4 SD=0.93; t=1.50, df = 38.51, p=0.14).

4.3 Production

Children in the child-human condition scored M=6.64 (SD=1.43) out of 9 on the production quiz and M=6.18 (SD=2.18) in the child-robot condition. Using a Welch Two Sample t-test no significant difference between the two conditions was found (t=-0.58, df =17.27, p=0.57).

We also analysed the total number of spatial vocabulary used in L2 (Figure 4). Due to a break in protocol, children were sometimes prompted to attempt a question again instead of moving on in the production quiz. As such our analysis is on words used without being prompted for an additional attempt. In the Robot condition, the children averaged M=9.45 (SD=2.46) spatial words, compared to M=9.36 (SD=1.91) in the Human condition. Using a Welch Two Sample t-test no significant difference was found (t=0.10, df=18.4, p=0.92).

Finally we analysed the amount and level of encouragement given (see levels in Section 3.3). While encoding encouragement given to the children we added a fourth level for analysis of the results:

(4) Encouragement is given that changes or disrupts the task, e.g. telling the child that the current question is the same as a previous one.

The mean amount of encouragement given was M=12.36 (SD=7.46) in the Human condition and M=13.09 (SD=7.78) in the Robot condition. No significant difference was found between the conditions (p=0.83). However we see a significant difference in the average maximum level of encouragement per question across the two conditions (Robot: M=1.12, SD=0.57. Adult: M=2.09, SD=1.09, p=0.02). This is strongly influenced by the amount of level 4 encouragement given by the adult, of which we see 33 instances across 10 children. We see a significant difference between the average amount of level 4 encouragement given per child between the amount given in the first half of the study compared to the second showing an increase in deviation from the protocol over time (First Half: M=1.25, SD=.0.88. Second Half: M=4.25, SD=2.64, p=0.04).

4.4 Comprehension and Production

The data we collected also provided us with an opportunity to test the predictions of Laufer et al. [9], a key foundation for our research.

By looking at the children's scores on comprehension (passive recognition) and production (active recall) we should be able to see evidence of a hierarchy, where comprehension is required for production.

Across both conditions the children had an average score on the production quiz of 6.41 (SD=1.82) out of 9 and is significantly above chance (p=0.03). A positive but non-significant correlation was found between the comprehension test score and their production quiz score (Pearson's r=0.29, p=0.19). The lack of a significant correlation suggests that abilities in comprehension and production are not directly related.

We marked a child as having achieved comprehension on a particular word if they required less than four attempts across the two relevant questions in the comprehension test. For example if we were looking at whether a child could comprehend the word 'sur' we would look at the number of attempts they took for questions three and five. If a child takes two attempts on question three and one attempt on question five their total number of attempts for 'sur' would be three. We would mark this child as being able to comprehend 'sur'. We marked a child as being able to produce a word if they scored at least two points in the production quiz on the three relevant questions. Using Guttman's Coefficient of Reproducibility (reported in Table 1), we were unable to find a hierarchy. A hierarchy would show that comprehension is needed for production. Guttman's Coefficient measures whether such a hierarchy exists based on the number of deviations from that hierarchy. A coefficient of over 0.9 is expected to display such a hierarchy.

	Sur	Sous	Devant
No. Deviations	5	3	4
Guttman's Coefficient λ_4	0.11	0.57	0.56

Table 1: Table detailing the number of deviations from the expected hierarchy and the Guttman's Coefficient of reproducibility. In the case of all three words, we fail to meet the reliability expectation of 0.9

5 DISCUSSION

5.1 Effectiveness of the robot to support L2 production

While this study does not show statistical improvement to a child's ability to produce by using a robot over a person, it does show a similar performance in this task, with no significant difference between the two conditions being found. It may still be desirable to use a robot to allow standardization and automation of assessment. With a minimal amount of support being provided by an agent, only a narrow set of phrases can be given – otherwise the nature of the task could be changed from production. This can make interactions very repetitive for the assessor. Though the scores were higher than expected it still proved to be a challenging task for the children. With the minimal amount of support available to an experimenter it could be emotionally stressful to be unable to intervene when a child is finding the task difficult.

The scores from the production quiz are higher than we expected. From the literature we expected L2 production to be difficult for the children, and our expert tutor believed that it would take two to three sessions for most children to produce at all. The observed prowess of the children may be partially explained by the design of the lessons, directly aimed at encouraging the children to produce the target words for this study. It should be noted that most productions were only single words. Only two children produced any of the support words (*nounours* – teddy bear, and *chaise* – chair).

Several factors may contribute to the high performance of the experimenter. Even within the context of a limited set of responses a person is able to provide much better cues and encouragement based on reading the child. These kind of social skills are still a gold standard to which robotics researchers strive. Though this experiment was conducted using a 'wizard', their position and the time delay in actions for the robot prevented this fine grained social interaction. Some of the cues provided by the experimenter were not programmed into the robot but should be added into its repertoire

(1) *Direct phonetic cues* - Giving part of the word e.g. the starting s.
(2) *Indirect phonetic cues* - Giving clues to the word about how it sounds e.g. "It's the one with a strange sound in it"
(3) *Rhythmic cues* - Giving the syllables of the word e.g. "Duh-dum". This may work well for the small target vocabulary, like ours, where this could refer to a single word, but may be less effective in larger vocabularies.
(4) *Gestural cues* - Movements with the hands that mimic gestures used by the teacher in the lesson.

Despite the more limited social skills of this implementation of the robot, it still achieved a similar performance level to a person. This may be the expected reduction of anxiety, that previous research has shown, balancing the limited social behaviours.

However we also saw a large amount of encouragement given to the children by the blind experimenter that was outside of the original protocol, that could be deemed to have affected the scores of the children in an undesirable way. While in the first half the amount of these encouragements by the experimenter remained low, there was a sharp increase in the latter half. This could be caused by forgetting the protocol over the days of the study or just growing more lax in its use, or even the emotional stress that is put on a person by the children's difficulties.

The presence of a wizard in the room may also have been a contributing factor. The presence of a person, even when not in view, may have prevented the robot from reducing anxiety as much as it could have done, as the child might be aware someone else is listening in. We minimized the affect of the wizard by ensuring there was no reason for them to interact with the children either before the study. Analysis of the videos showed that the majority of children never turned towards the wizard at any point during the study, and focused on the robot. So we believe the impact of the wizard's presence was minimal.

Finally, it must be noted that the school where we performed the study cultivated a much friendlier relationship between adults in the school and the students than is typically seen. This may have made the children feel more comfortable and confident in the presence of our experimenter, reducing anxiety. Future work will

focus on broadening this study to multiple schools to see whether our results can be replicated in different settings.

5.2 Relative difficulty of comprehension versus production

The lack of correlation shown between the production quiz score and the number of attempts on the comprehension test (Figure 6) shows that there was no direct relation between comprehension and production vocabularies. However when we look at the possibility of a hierarchy from comprehension to production we do not find evidence to support a hierarchy. This could have had several causes. While we were hoping to find support within our data, we were not directly testing for this hierarchy. Laufer et al. [9] looked at students 16 years and older at high school and university who had been studying their L2 as part of a national curriculum for between 6 to 9 years. Ours is based on a single lesson focused entirely on being able to say the target words. The younger children in our study may also have been more receptive to learning words productively, as they are still increasing their phonological vocabulary. These skills have been shown to have a correlation with word vocabulary [6]. These factors could account for an increase in deviations from the previously established hierarchy.

6 CONCLUSION

We hypothesized that a robot could surpass human performance in encouraging the production of spatial language: this hypothesis is not supported by our study; however, the robot and the facilitator's performance were very similar, with no significant difference between the two conditions being found. This was despite the greater social ability of the human experimenter. This may be explained by the previous research that shows that robots can make people less anxious in foreign language learning scenarios. Future work expanding the robot's social ability may improve the robot's ability to assess and support a student's learning.

Measuring the production skills of a child at this level is a repetitive and lengthy task. An autonomous robot that is able to measure the production level of a child could be used as a tool to alleviate these factors, enabling more accurate data collection for both research and assessment purposes. Currently we are planning on expanding this work to more schools while increasing the social skills of the robot.

7 ACKNOWLEDGEMENTS

This work was supported by the EU H2020 L2TOR project (grant 688014). The authors would also like to thank the teacher, who wished to remain anonymous, who provided the French lessons for the children. All statistics and graphs were obtained using R [15].

REFERENCES

[1] Minoo Alemi. 2016. General Impacts of Integrating Advanced and Modern Technologies on Teaching English as a Foreign Language. *International Journal on Integrating Technology in Education* 5, 1 (2016), 13–26.
[2] M Alemi, A Meghdari, and M Ghazisaedy. 2015. The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics* (2015), 1–13.
[3] Paul Baxter, Rachel Wood, and Tony Belpaeme. 2012. A touchscreen-based'sandtray'to facilitate, mediate and contextualise human-robot social interaction. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 105–106.
[4] Larry Fenson, Philip S. Dale, Steven Reznick, Donna J. Thal, Elizabeth Bates, Jeff Hartung, Stephen J. Pethick, and Judy Reilly. 1993. *The MacArthur Communicative Development Inventories: UserâĂŹs guide and technical manual.* San Diego, CA: Singular Publishing.
[5] Morrison F. Gardner. 1990. *Expressive One-Word Picture Vocabulary Test - Revised.* Novato, CA: Academic Therapy.
[6] Susan E Gathercole and Alan D Baddeley. 1989. Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of memory and language* 28, 2 (1989), 200–213.
[7] James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme. 2016. Heart vs hard drive: children learn more from a human tutor than a social robot. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 451–452.
[8] Batia Laufer. 1998. The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics* 19, 2 (1998), 255–271.
[9] Batia Laufer and Zahava Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning* 54, 3 (2004), 399–436.
[10] Batia Laufer and Paul Nation. 1999. A vocabulary-size test of controlled productive ability. *Language Testing* 16, 1 (1999), 33–51.
[11] Batia Laufer and T. Sima Paribakht. 1998. The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning* 48, 3 (1998), 365–391.
[12] Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, and Munsang Kim. 2011. On the effectiveness of robot-assisted language learning. *ReCALL* 23, 01 (2011), 25–58.
[13] Didier Maillat. 2010. The pragmatics of L2 in CLIL. Language use and language learning in CLIL classrooms. *Language Use and Language Learning in CLIL Classrooms* (2010), 39–58.
[14] Jan-Arjen Mondria and Boukje Wiersma. 2004. Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In *Vocabulary in a Second Language: Selection, Acquisition and Testing*, Paul Bogaards and Batia Laufer (Eds.). John Benjamins Publishers, 79–100.
[15] R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[16] Fumihide Tanaka and Shizuko Matsuzoe. 2012. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction* 1, 1 (2012).
[17] E. H. Wiig, W. Secord, and E. Semel. 1992. *CELF-Preschool: Clinical Evaluation of Language Fundamentals - Preschool.* New York: Psychological Corp.
[18] Kathleen Williams. 1997. *Expressive Vocabulary Test.* Minnesota: American Guidance Service.

Measuring Engagement Online in CRI

Pieter Wolfert
Tilburg University
Tilburg, The Netherlands
a.p.wolfert@uvt.nl

Paul Vogt
Tilburg University
Tilburg, The Netherlands
p.a.vogt@uvt.nl

Mirjam de Haas
Tilburg University
Tilburg, The Netherlands
mirjam.dehaas@uvt.nl

Pim Haselager
Radboud University
Nijmegen, the Netherlands
w.haselager@donders.ru.nl

The Near Future of Children's Robotics, IDC Workshop 2018, Trondheim, Norway

Abstract
In this paper, a pipeline is suggested for measuring child engagement in a robot tutoring task, together with a pilot experiment for verification. Smiling, gaze direction and posture are taken as indicators for engagement. A pilot experiment is proposed to test the performance of the model. This will be a robot tutoring task based on the child game "I spy with my little eye" during which children with the age of five learn English names for animals [6]. In this pilot experiment, the children are provided breaks when they are dis-engaged, to re-engage the children. Afterwards, the children will be asked to rate the perception of the robot, and it is expected that this rating will be higher than a robot without engagement detection.

Author Keywords
Child Robot Interaction; Machine Learning; Applied Robotics; Robot Tutoring

ACM Classification Keywords
· **Human-centered computing** → **HCI design and evaluation methods;** [· **Human-centered computing** → **Human computer interaction (HCI)**]

Introduction

In order to make tutor robots more effective, robots should be able to adapt to a user. In a study done by Leyzberg et al., [10] where students had to solve a grid-based logic puzzle, it was found that personalized tutoring improved the students' performance. Ramachandran et al. [12] found that providing personal breaks based on the number of errors during a task in a robot tutoring session improved learning in children. Improving the effectiveness of a robot tutoring task can be done in several ways. One way is to keep track of a child's progress over time, by registering the words that a child learnt [6, 12]. In a previous experiment of De Wit et al. [6] was found that adaptive tutoring did not have a significant effect on learning gain, but the use of gestures did have a significant effect. An important finding, however, was that engagement seemed to lower at the end of the interaction, in all conditions. With this in mind, another way is to improve the quality of the time a child spents with the robot. This can be done by adjusting the robots behavior such that it creates a better environment for interaction and tutoring, based on the child's engagement [14].

A robot that can detect a child's (lack of) engagement online during the interaction and respond to this appropriately could improve the quality of tutoring by re-engaging the child when so desired. It has been found that eye-gaze is a good indicator of engagement [8], and eye-gaze is proposed as a metric for engagement in human robot interaction [15, 1]. Body posture and head position are also identified as indicators for engagement [1], head pose is also used for realtime attention assessment [9]. Another predictor for engagement is smiling [5, 15]. Serholt et al. [15] found that smiles were most common after positive feedback. Measuring smiles can be done by using a multilayer convolutional neural network trained on facial expressions [2].

In this paper, a pipeline is proposed to measure engagement based on smiling, posture and gaze of a child during a robot tutoring task; such that an online prediction can be generated given a frontal video recording. In order to verify whether the pipeline works, a pilot experiment is suggested based on the experiment done by de Wit et al. [6].

This project is embedded in the L2TOR project. The L2TOR project ('el tutor') aims to design a social robot that supports the teaching of a second language to preschool children. This platform, which runs on a NAO humanoid robot, requires that the robot is able to work together with young children (age 5), on a peer level, and can provide relevant feedback.

Methods

The aim of this project is to construct and train a model to do online engagement prediction. This model will consist of three different neural networks (one for each feature), with a combined output resulting in a prediction. Furthermore, a method is developed to re-engage the child when the child seems disengaged. This method will then be tested and verified in a pilot experiment based on de Wit et al. [6], in a NAO humanoid robot. In the current L2TOR experiments the robot does not account for the child's engagement. By accounting for the engagement of a child with the robot, the behavior of the robot can be more matched to the child's behavioral state, by for example providing a break when the engagement seems to drop. Engagement will be measured by combining smiling, posture and gaze predictions. Two datasets will be used for engagement prediction: a dataset of a previous L2TOR study [6] and the EmoReact dataset [11]. The L2TOR dataset contains frontal video clips of subjects performing a language learning task with the robot. This dataset has been annotated with engagement observations, in the range low, medium, high. In each

frame the face and the upper body of the child is visible.

The EmoReact dataset contains images annotated with emotional states of children between the ages 4 and 14. There are in total 17 states annotated. A subset will be used such that only 4 and 5 year old children are considered. The L2TOR dataset provides the engagement observations and the EmoReact dataset images with emotional expressions of children, useful for detecting smiles. The smiling prediction will be done using a convolutional neural network based on the model proposed by Arriaga et al. [2].

OpenPose [4] will be used to extract postures from the L2TOR dataset, this information is then, together with the engagement labels, used to train a model that can classify engagement. In order to do gaze prediction, the L2TOR data will be processed to identify both the position of the robot and the tablet (the tablet is part of the tutoring experiments). To perform gaze prediction a pretrained model provided by Recasens et al. [13] is used. This model provides the gaze direction of a child given the location of the face of the child in the video. Localisation of the face of the child will be done using OpenCV [3]. The three different neural networks' outputs are followed by an LSTM layer [7] –to make sure that temporal relations are learned– necessary for measuring engagement. Once this model is trained, it can be used online during an experiment and provide engagement predictions.

In the pilot experiment to evaluate the effectiveness of the model, participants (children) will learn six English animal words during a game of "I spy with my little eye". During the task a NAO humanoid robot will provide an animal name in English, after which the child has to select the correct animal on a tablet. In this pilot experiment, the robot will offer breaks during the experiment based on the engagement of the child, after this break the session continues. We expect that the engagement will increase after the break, based on the experiment of Ramachandran [12].

Conclusion

In this paper a model for online measuring of children's engagement in a robot tutoring task based on smiling, gaze and posture is proposed. A valid model for online measuring of engagement is not available yet, while such a model could influence learning in robot tutoring and can stimulate the developments of robot tutoring systems. The model and methods can be adjusted to work with other domains as well, outside the scope of robot tutoring, enhancing measuring of engagement in human robot interaction. The proposed model is currently being developed, and the expectation is that preliminary results of both the model and the evaluation pilot will be presented at the CRI-IDC Workshop.

Acknowledgments

This work has been supported by the EU H2020 L2TOR project (grant 688014).

REFERENCES

1. Salvatore M Anzalone, Sofiane Boucenna, Serena Ivaldi, and Mohamed Chetouani. 2015. Evaluating the engagement with social robots. *International Journal of Social Robotics* 7, 4 (2015), 465–478.
2. Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. 2017. Real-time Convolutional Neural Networks for Emotion and Gender Classification. *arXiv preprint arXiv:1710.07557* (2017).
3. Gary Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).
4. Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, Vol. 1. 7.

5. Ginevra Castellano, André Pereira, Iolanda Leite, Ana Paiva, and Peter W McOwan. 2009. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 119–126.
6. Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 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*. ACM, 50–58.
7. Sepp Hochreiter and Jurgen Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1–32.
8. Ryo Ishii, Yukiko I Nakano, and Toyoaki Nishida. 2013. Gaze awareness in conversational agents: Estimating a user's conversational engagement from eye gaze. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 2 (2013), 11.
9. Séverin Lemaignan, Fernando Garcia, Alexis Jacq, and Pierre Dillenbourg. 2016. From real-time attention assessment to with-me-ness in human-robot interaction. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 157–164.
10. Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. 2014. Personalizing robot tutors to individuals' learning differences. *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14* March 2014 (2014), 423–430.
11. Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. 2016. EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 137–144.
12. Aditi Ramachandran, C.-M. Huang, and Brian Scassellati. 2017. Give Me a Break!: Personalized Timing Strategies to Promote Learning in Robot-Child Tutoring. *ACM/IEEE International Conference on Human-Robot Interaction* Part F1271 (2017), 146–155.
13. Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Advances in Neural Information Processing Systems*. 199–207.
14. Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. 2017. Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17* (2017), 128–136.
15. Sofia Serholt and Wolmet Barendregt. 2016. Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction. *ACM International Conference Proceeding Series* 23-27-Octo (2016).

Playful exploration of a robot's gesture production and recognition abilities

Jan de Wit[1], Bram Willemsen[1], Mirjam de Haas[1], Pieter Wolfert[1,2], Paul Vogt[1], Emiel Krahmer[1]

[1] Tilburg University [2] Radboud University Nijmegen

A robot's ability to communicate non-verbally can increase understanding between human and robot, and can help to maintain an engaging interaction. However, gestures in most studies with robots tend to rely on the designers' frame of reference, and their perspective on the robot's physical limitations. We propose a system, based on a gesture guessing game, where the robot learns many different examples of gestures from human players, which it can then replicate during subsequent interactions. Because players attempt to guess the meaning of the robot's gestures, the robot is able to identify those examples that best represent the target objects, given its limited expressive abilities. A first iteration of this exploratory study is set to take place with a SoftBank NAO robot, at the NEMO science museum in Amsterdam.

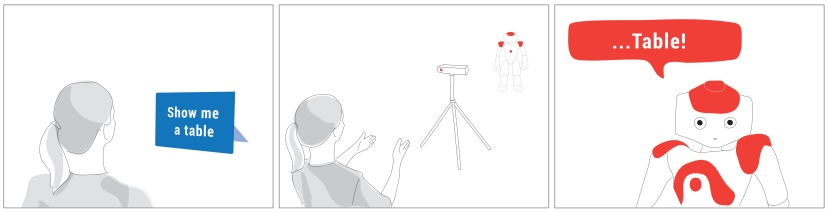


Figure 1: Example of the robot's turn to guess.

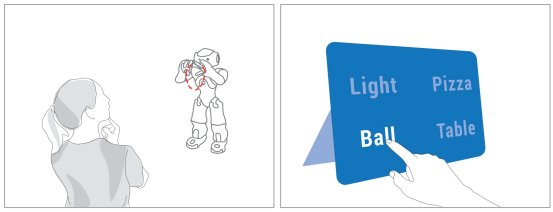


Figure 2: Example of the player's turn to guess.

Feature extraction

To detect similarity between gestures regardless of variations in the speed or size of the motion, we extract the *gist* of the gesture, based on the inflection points in the motion's trajectory [1]. We then find the nearest extreme point of the hand's position (Figure 3) and map this to a quadrant that is relative to the person's shoulder (Figure 4). This results in a description of a gesture that consists of a sequence of salient points from the motion's trajectory.

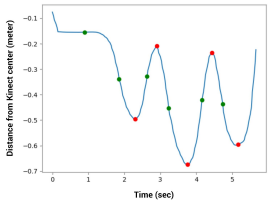


Figure 3: Inflection points (green) and peaks (red) of a motion.

Gesture recognition

Gesture recognition is currently implemented with a k-nearest neighbors approach, where the similarity between gestures is measured by aligning them using the Needleman-Wunsch algorithm with a custom scoring matrix.

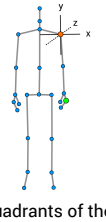


Figure 4: Quadrants of the hand position relative to the shoulder.

Gesture generation

Recorded examples of gestures for each object are clustered based on their similarity. Clusters may thus represent different strategies or variations within the gestures. The robot picks the next gesture to perform based on the weights assigned to each cluster and to individual examples within the cluster (through exploration and exploitation). Depending on whether the player guesses the object correctly or not, the weights are either increased or decreased.

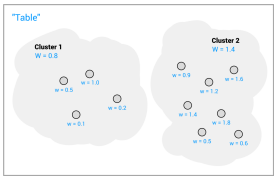


Figure 5: Example of clustered gestures with weights.

[1] Maria Eugenia Cabrera and Juan Wachs. A Human-Centered approach to One-Shot Gesture Learning. Frontiers in Robotics and AI 4 (2017).

PLOS ONE

RESEARCH ARTICLE

The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics

Séverin Lemaignan[1]*, **Charlotte E. R. Edmunds**[2], **Emmanuel Senft**[2], **Tony Belpaeme**[2,3]

1 Bristol Robotics Lab, University of the West of England, Bristol, United Kingdom, **2** Centre for Robotics and Neural Systems, University of Plymouth, Plymouth, United Kingdom, **3** IDLab – imec, Ghent University, Ghent, Belgium

* severin.lemaignan@brl.ac.uk

Citation: Lemaignan S, Edmunds CER, Senft E, Belpaeme T (2018) The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. PLoS ONE 13(10): e0205999. https://doi.org/10.1371/journal.pone.0205999

Editor: Michael L. Goodman, University of Texas Medical Branch at Galveston, UNITED STATES

Received: June 5, 2018

Accepted: October 4, 2018

Published: October 19, 2018

Data Availability Statement: The dataset is freely available to any interested researcher. Due to ethical and data protection regulations, the dataset is however made available in two forms: - a public, Creative Commons licensed, version that does not include any video material of the children (no video nor audio streams), and hosted on the Zenodo open-data platform: https://zenodo.org/record/1043508. - the complete version that includes all video streams is freely available as well, but interested researchers must first fill a data protection form. The detail of the procedure are

Abstract

The study of the fine-grained social dynamics between children is a methodological challenge, yet a good understanding of how social interaction between children unfolds is important not only to Developmental and Social Psychology, but recently has become relevant to the neighbouring field of Human-Robot Interaction (HRI). Indeed, child-robot interactions are increasingly being explored in domains which require longer-term interactions, such as healthcare and education. For a robot to behave in an appropriate manner over longer time scales, its behaviours have to be contingent and meaningful to the unfolding relationship. Recognising, interpreting and generating sustained and engaging social behaviours is as such an important—and essentially, open—research question. We believe that the recent progress of machine learning opens new opportunities in terms of both analysis and synthesis of complex social dynamics. To support these approaches, we introduce in this article a novel, open dataset of child social interactions, designed with data-driven research methodologies in mind. Our data acquisition methodology relies on an engaging, methodologically sound, but purposefully underspecified *free-play* interaction. By doing so, we capture a rich set of behavioural patterns occurring in natural social interactions between children. The resulting dataset, called the PInSoRo dataset, comprises 45+ hours of hand-coded recordings of social interactions between 45 child-child pairs and 30 child-robot pairs. In addition to annotations of social constructs, the dataset includes fully calibrated video recordings, 3D recordings of the faces, skeletal informations, full audio recordings, as well as game interactions.

Introduction

Studying social interactions

Studying social interactions requires a social *situation* that effectively elicits interactions between the participants. Such a situation is typically scaffolded by a social task, and consequently, the nature of this task influences in fundamental ways the kind of interactions that

available online: https://freeplay-sandbox.github.io/application.

Funding: This work was primarily funded by the European Union H2020 "Donating Robots a Theory of Mind" project (grant id #657227) awarded to SL. It received additional funding from the European Union H2020 "Second Language Tutoring using Social Robots" project (grant id #688014), awarded to TB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

might be observed and analysed. In particular, the socio-cognitive tasks commonly found in both the experimental psychology and human-robot interaction (HRI) literature often have a narrow focus: because they aim at studying one (or a few) specific social or cognitive skills in isolation and in a controlled manner, these tasks are typically conceptually simple and highly constrained (for instance, object hand-over tasks; perspective-taking tasks; etc.). While these focused endeavours are important and necessary, they do not adequately reflect the complexity and dynamics of real-world, natural interactions (as discussed by Baxter et al. in [1], in the context of HRI). Consequently, we need to investigate richer interactions, scaffolded by socio-cognitive tasks that:

- are long enough and varied enough to elicit a large range of interaction situations;
- foster rich multi-modal interactions, such as simultaneous speech, gesture, and gaze behaviours;
- are not over-specified, in order to maximise natural, non-contrived behaviours;
- evidence complex social dynamics, such as rhythmic coupling, joint attention, implicit turn-taking;
- include a level of non-determinism and unpredictability.

The challenge lies in designing a social task that exhibits these features *while maintaining* essential scientific properties (repeatability; replicability; robust metrics) as well as good practical properties (not requiring unique or otherwise very costly experimental environments; not requiring very specific hardware or robotic platform; easy deployment; short enough experimental sessions to allow for large groups of participants).

Looking specifically at social interactions amongst children, we present in the next section our take on this challenge, and we introduce a novel task of free play. The task is designed to elicit rich, complex, varied social interactions while supporting rigorous scientific methodologies, and is well suited for studying both child-child and child-robot interactions.

Social play

Our interaction paradigm is based on free and playful interactions (hereafter, *free play*) in what we call a *sandboxed environment*. In other words, while the interaction is free (participants are not directed to perform any particular task beyond playing), the activity is both *scaffolded* and *constrained* by the setup mediating the interaction (a large interactive table), in a similar way to children freely playing with sand within the boundaries of a sandpit. Consequently, while participants engage in open-ended and non-directed activity, the play situation is framed to be easily reproducible as well as practical to record and analyse.

This initial description frames the socio-cognitive interactions that might be observed and studied: playful, dyadic, face-to-face interactions. While gestures and manipulations (including joint manipulations) play an important role in this paradigm, the participants do not typically move much during the interaction. Because it builds on play, this paradigm is also primarily targeted to practitioners in the field of child-child or child-robot social interactions.

The choice of a playful interaction is supported by the wealth of social situations and social behaviours that play elicits (see for instance parts 3 and 4 of [2]). Most of the research in this field builds on the early work of Parten who established five *stages of play* [3], corresponding to different stages of development, and accordingly associated with typical age ranges: (*a*) *solitary (independent) play* (age 2-3): child playing separately from others, with no reference to what others are doing; (*b*) *onlooker play* (age 2.5-3.5): child watching others play; may engage

in conversation but not engage in doing; true focus on the children at play; (*c*) *parallel play* (also called adjacent play, social co-action, age 2.5-3.5): children playing with similar objects, clearly beside others but not with them; (*d*) *associative play* (age 3-4): child playing with others without organization of play activity; initiating or responding to interaction with peers; (*e*) *cooperative play* (age 4+): coordinating one's behavior with that of a peer; everyone has a role, with the emergence of a sense of belonging to a group; beginning of "team work."

These five stages of play have been extensively discussed and refined over the last century, yet remain remarkably widely accepted. It must be noted that the age ranges are only indicative. In particular, most of the early behaviours still occur at times by older children.

Machine learning, robots and social behaviours

The data-driven study of social mechanisms is still an emerging field, and only limited literature is available.

The use of interaction datasets to teach artificial agents (robots) how to socially behave has been previously explored, and can be considered as the extension of the traditional learning from demonstration (LfD) paradigms to social interactions [4, 5]. However, existing research focuses on low-level identification or generation of brief, isolated behaviours, including social gestures [6] and gazing behaviours [7].

Based on a human-human interaction dataset, Liu et al. [8] have investigated machine learning approaches to learn longer interaction sequences. Using unsupervised learning, they train a robot to act as a shop-keeper, generating both speech and socially acceptable motions. Their approach remains task-specific, and they report only limited success. They however emphasise the "life-likeness" of the generated behaviours.

This burgeoning interest in the research community for the data-driven study of social responses is however impaired by the lack of structured research efforts. In particular, there is only limited availability of large and open datasets of social interactions, suitable for machine-learning applications.

One such dataset is the *Multimodal Dyadic Behavior Dataset* (*MMDB*, [9]). It comprises of 160 sessions of 3 to 5 minute child-adult interactions. During these interactions, the experimenter plays with toddlers (1.5 to 2.5 years old) in a semi-structured manner. The dataset includes video streams of the faces and the room, audio, physiological data (electrodermal activity) as well as manual annotations of specific behaviours (like gaze to the examiner, laughter, pointing). This dataset focuses on very young children during short, adult-driven interactions. As such, it does not include episodes of naturally-occurring social interactions between peers, and the diversity of said interactions is limited. Besides, the lack of intrinsic and extrinsic camera calibration information in the dataset prevent the automatic extraction and labeling of key interaction features (like mutual gaze).

Another recent dataset, the *Tower Game Dataset* [10], focuses specifically on rich dyadic social interactions. The dataset comprises of 39 adults recorded over a total of 112 annotated sessions of 3 min in average. The participants are instructed to jointly construct a tower using wooden blocks. Interestingly, the participants are not allowed to talk to maximise the amount of non-verbal communication. The skeletons and faces of the participants are recorded, and the dataset is manually annotated with so-called *Essential Social Interaction Predicates* (ESIPs): rhythmic coupling (entrainment or attunement), mimicry (behavioral matching), movement simultaneity, kinematic turn taking patterns, joint attention. This dataset does not appear to be publicly available on-line.

The UE-HRI dataset [11] is another recently published (2017) dataset of social interactions, focusing solely on human-robot interactions. 54 adult participants were recorded (duration

M = 7.7min) during spontaneous dialogues with a Pepper robot. The interactions took place in a public space, and include both one-to-one and multi-party interactions. The resulting dataset includes audio and video recordings from the robot perspective, as well as manual annotations of the levels of engagement. It is publicly available.

PInSoRo, our dataset, shares some of the aims of the *Tower Game* and *UE-HRI* datasets, with however significant differences. Contrary to these two datasets, our target population are children. We also put a strong focus on naturally occurring, real-world social behaviours. Furthermore, as presented in the following sections, we record much longer interactions (up to 40 minutes) of free play interactions, capturing a wider range of socio-cognitive behaviours. We did not place any constraints on the permissible communication modalities, and the recordings were manually annotated with a focus on social constructs.

Material and methods

The free-play sandbox task

As previously introduced, the *free-play sandbox* task is based on face-to-face free-play interactions, mediated by a large, horizontal touchscreen. Pairs of children (or alternatively, one child and one robot) are invited to freely draw and interact with items displayed on an interactive table, without any explicit goals set by the experimenter (Fig 1). The task is designed so that children can engage in open-ended and non-directive play. Yet, it is sufficiently constrained to

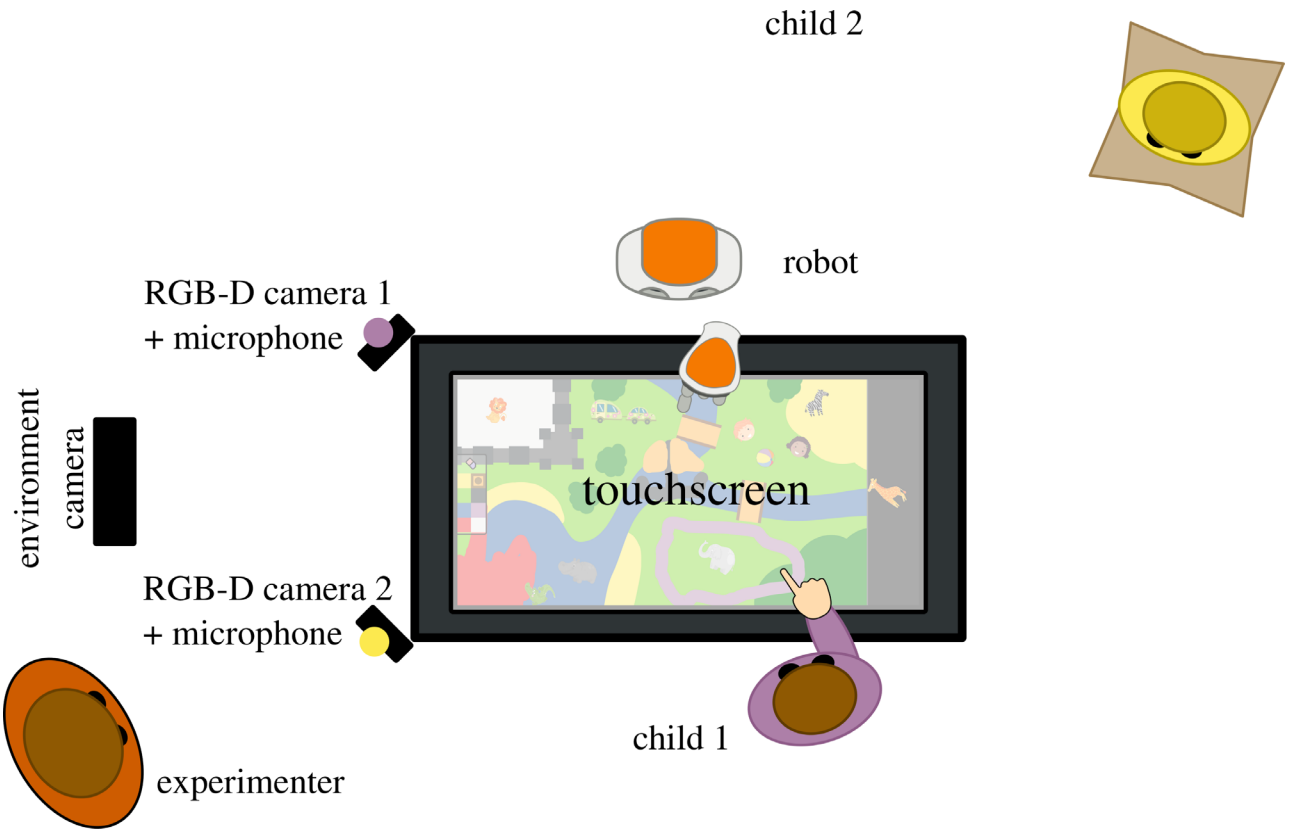


Fig 1. The free-play social interactions sandbox: Two children or one child and one robot (as pictured here) interacted in a free-play situation, by drawing and manipulating items on a touchscreen. Children were facing each other and sit on cushions. Each child wore a bright sports bib, either purple or yellow, to facilitate later identification.

https://doi.org/10.1371/journal.pone.0205999.g001

Fig 2. Example of a possible game situation. Game items (animals, characters. . .) can be dragged over the whole play area, while the background picture can be painted over by picking a colour. In this example, the top player is played by a robot.

https://doi.org/10.1371/journal.pone.0205999.g002

be suitable for recording, and allows the reproduction of social behaviour by an artificial agent in comparable conditions.

Specifically, the free-play sandbox follows the *sandtray* paradigm [12]: a large touchscreen (60cm × 33cm, with multitouch support) is used as an interactive surface. The two players, facing each other, play together, moving interactive items or drawing on the surface if they wish so (Fig 2). The background image depicts a generic empty environment, with different symbolic colours (water, grass, beach, bushes. . .). By drawing on top of the background picture, the children can change the environment to their liking. The players do not have any particular task to complete, they are simply invited to freely play. They can play for as long as they wish. However, for practical reasons, we had to limit the sessions to a maximum of 40 minutes.

Even though the children do typically move a little, the task is fundamentally a face-to-face, spatially delimited, interaction, and as such simplifies the data collection. In fact, the children's faces were successfully detected in 98% of the over 2 million frames recorded during the PInSoRo dataset acquisition campaign.

Experimental conditions. The PInSoRo dataset aims to establish two experimental baselines for the free-play sandbox task: the 'human social interactions' baseline on one hand (child–child condition), an 'asocial' baseline on the other hand (child–*non-social* robot condition). These two baselines aim to characterise the qualitative and quantitative bounds of the spectrum of social interactions and dynamics that can be observed in this situation.

In the *child-child* condition, a diverse set of social interactions and social dynamics were expected to be observed, ranging from little social interactions (for instance, with shy children) to strong, positive interactions (for instance, good friends), to hostility (children who do not get along very well).

In the *asocial* condition, one child was replaced by an autonomous robot. The robot was purposefully programmed to be *asocial*. It autonomously played with the game items as a child would (although it did not perform any drawing action), but avoided all social interactions: no social gaze, no verbal interaction, no reaction to child-initiated game actions.

From the perspective of social psychology, this condition provides a baseline for the social interactions and dynamics at play (or the lack thereof) when the social communication channel is severed between the agents, while maintaining a similar social setting (face-to-face interaction; free-play activity).

From the perspective of human-robot interaction and artificial intelligence in general, the child–'asocial robot' condition provides a baseline to contrast with for yet-to-be-created richer social and behavioural AI policies.

Hardware apparatus. The interactive table was based on a 27" Samsung All-In-One computer (quad core i7-3770T, 8GB RAM) running Ubuntu Linux and equipped with a fast 1TB SSD hard-drive. The computer was held horizontally in a custom aluminium frame standing 26cm above the floor. All the cameras were connected to the computer via USB-3. The computer performed all the data acquisition using ROS Kinetic (http://www.ros.org/). The same computer was also running the game interface on its touch-enabled screen (60cm × 33cm), making the whole system standalone and easy to deploy.

The children's faces were recorded using two short range (0.2m to 1.2m) Intel RealSense SR300 RGB-D cameras placed at the corners of the touchscreen (Fig 1) and tilted to face the children. The cameras were rigidly mounted on custom 3D-printed brackets. This enabled a precise measurement of their 6D pose relative to the touchscreen (extrinsic calibration).

Audio was recorded from the same SR300 cameras (one mono audio stream was recorded for each child, from the camera facing him or her).

Finally, a third RGB camera (the RGB stream of a Microsoft Kinect One, the *environment camera* in Fig 1) recorded the whole interaction setting. This third video stream was intended to support human coders while annotating the interaction, and was not precisely calibrated.

In the child-robot condition, a Softbank Robotics' Nao robot was used. The robot remained in standing position during the entire play interaction. The actual starting position of the robot with respect to the interactive table was recalibrated before each session by flashing a 2D fiducial marker on the touchscreen, from which the robot could compute its physical location.

Software apparatus. The software-side of the free-play sandbox is entirely open-source (source code: https://github.com/freeplay-sandbox/). It was implemented using two main frameworks: Qt QML (http://doc.qt.io/qt-5/qtquick-index.html) for the user interface (UI) of the game (Fig 2), and the *Robot Operating System* (ROS) for the modular implementation of the data processing and behaviour generation pipelines, as well as for the recordings of the various datastreams (Fig 4). The graphical interface interacts with the decisional pipeline over a bidirectional QML-ROS bridge that was developed for that purpose (source code available from the same link).

Fig 3 presents the complete software architecture of the sandbox as used in the child-robot condition (in the child-child condition, robot-related modules were simply not started).

Robot control. As previously described, one child was replaced by a robot in the child-robot condition. Our software stack allowed for the robot to be used in two modes of operations: either autonomous (selecting actions based on pre-programmed play policies), or controlled by a human operator (so-called *Wizard-of-Oz* mode of operation).

For the purpose of the PInSoRo dataset, the robot behaviour was fully autonomous, yet coded to be purposefully *asocial* (no social gaze, no verbal interaction, no reaction to child-initiated game actions). The simple action policy that we implemented consisted in the robot choosing a random game item (in its reach), and moving that item to a predefined zone on the

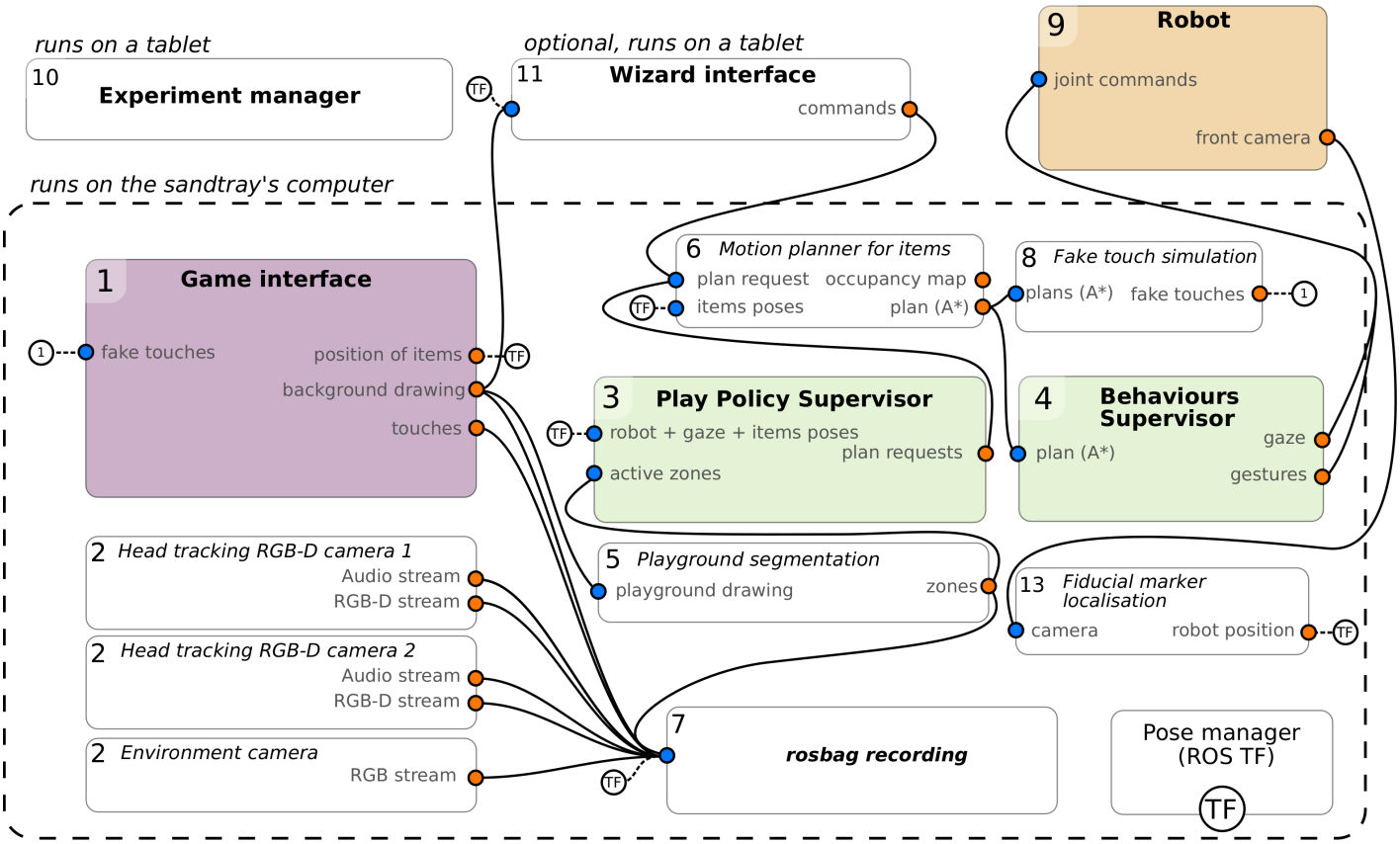


Fig 3. Software architecture of the free-play sandbox (data flows *from* orange dots *to* blue dots). Left nodes interact with the interactive table hardware (game interface (1) and camera drivers (2)). The green nodes in the centre implement the behaviour of the robot (play policy (3) and robot behaviours (4)). Several helper nodes are available to provide for instance a segmentation of the children drawings into zones (5) or A* motion planning for the robot to move in-game items (6). Nodes are implemented in Python (except for the game interface, developed in QML) and inter-process communication relies on ROS. 6D poses are managed and exchanged via ROS TF.

https://doi.org/10.1371/journal.pone.0205999.g003

map (e.g. if the robot could reach the crocodile figure, it would attempt to drag it to a blue, i.e. water, zone). The robot did not physically drag the item on the touchscreen: it relied on a A* motion planner to find an adequate path, sent the resulting path to the touchscreen GUI to animate the displacement of the item, and moved its arm in a synchronized fashion using the inverse kinematics solver provided with the robot's software development kit (SDK).

In the Wizard-of-Oz mode of operation, the experimenter would remotely control the robot through a tablet application developed for this purpose (Figs 3–11). The tablet exactly mirrored the game state, and the experimenter dragged the game items on the tablet as would the child on the touchscreen. On release, the robot would again mimic the dragging motion on the touchscreen, moving an object to a new location. This mode of operation, while useful to conduct controlled studies, was not used for the dataset acquisition.

Experiment manager. We developed as well a dedicated web-based interface (usually accessed from a tablet) for the experimenter to manage the whole experiment and data acquisition procedure (Figs 3–10). This interface ensured that all the required software modules were running; it allowed the experimenter to check the status of each of them and, if needed, to start/stop/restart any of them. It also helped managing the data collection campaign by

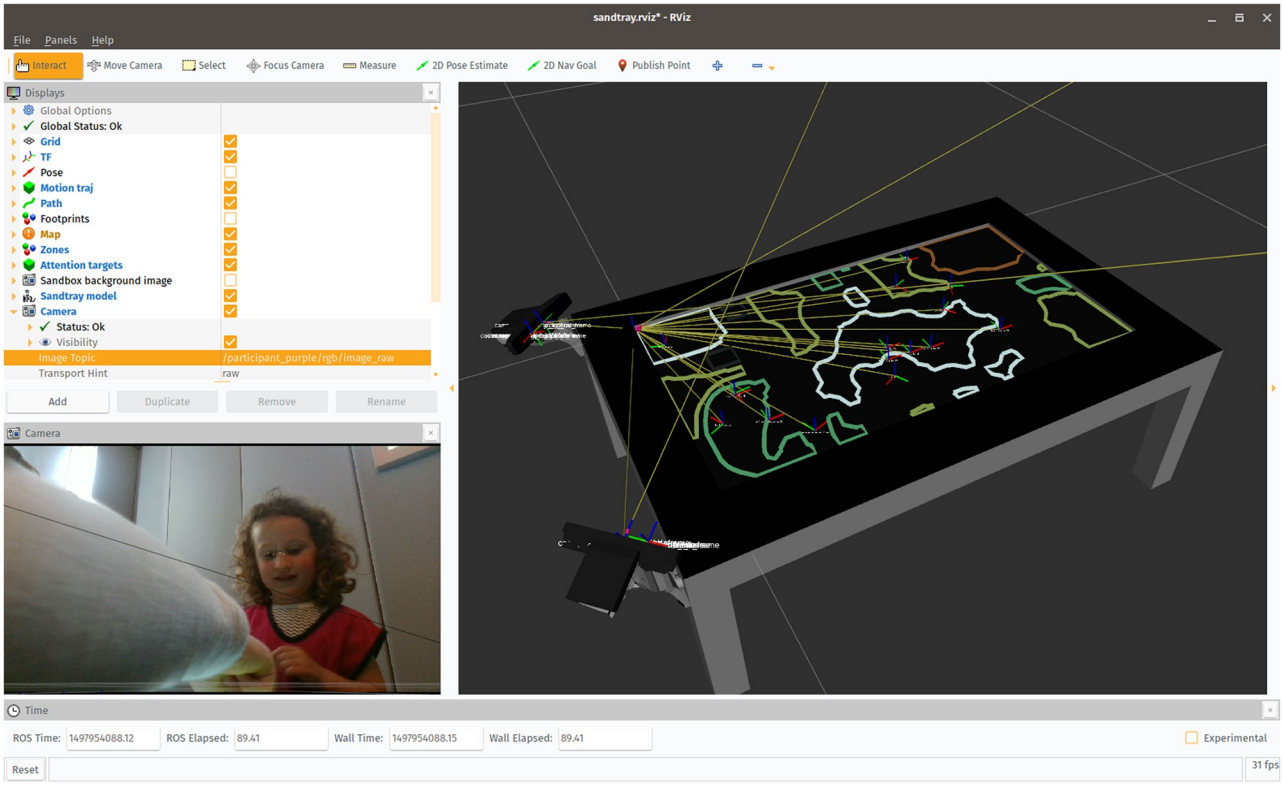


Fig 4. The free-play sandbox, viewed at runtime within ROS RViz. Simple computer vision was used to segment the background drawings into zones (visible on the right panel). The poses and bounding boxes of the interactive items were broadcast as well, and turned into an occupancy map, used to plan the robot's arm motion. The individual pictured in this figure has given written informed consent (as outlined in PLOS consent form) to appear.

https://doi.org/10.1371/journal.pone.0205999.g004

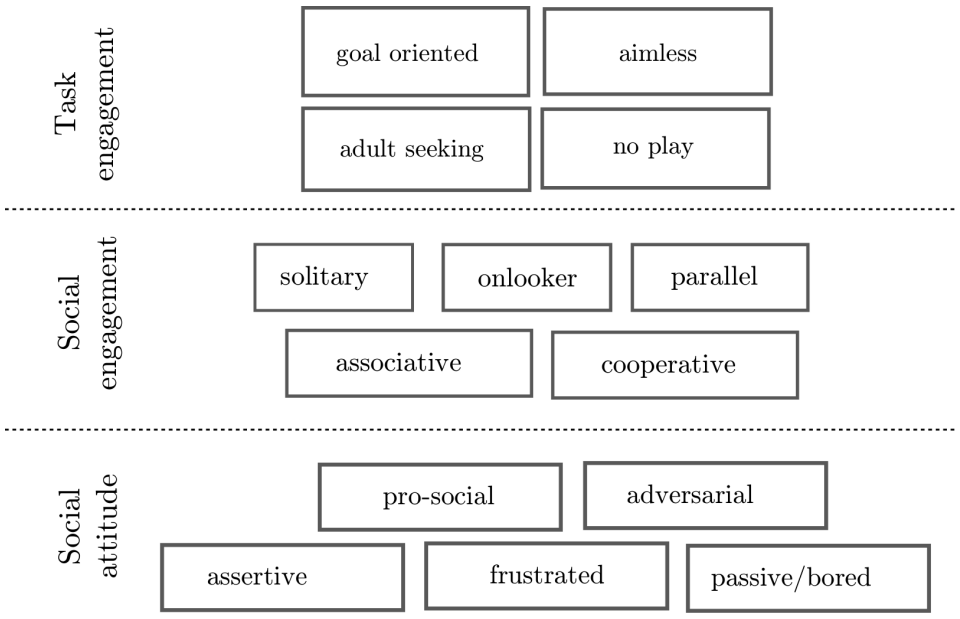


Fig 5. The coding scheme used for annotating social interactions occurring during free-play episodes. Three main axis were studied: task engagement, social engagement and social attitude.

https://doi.org/10.1371/journal.pone.0205999.g005

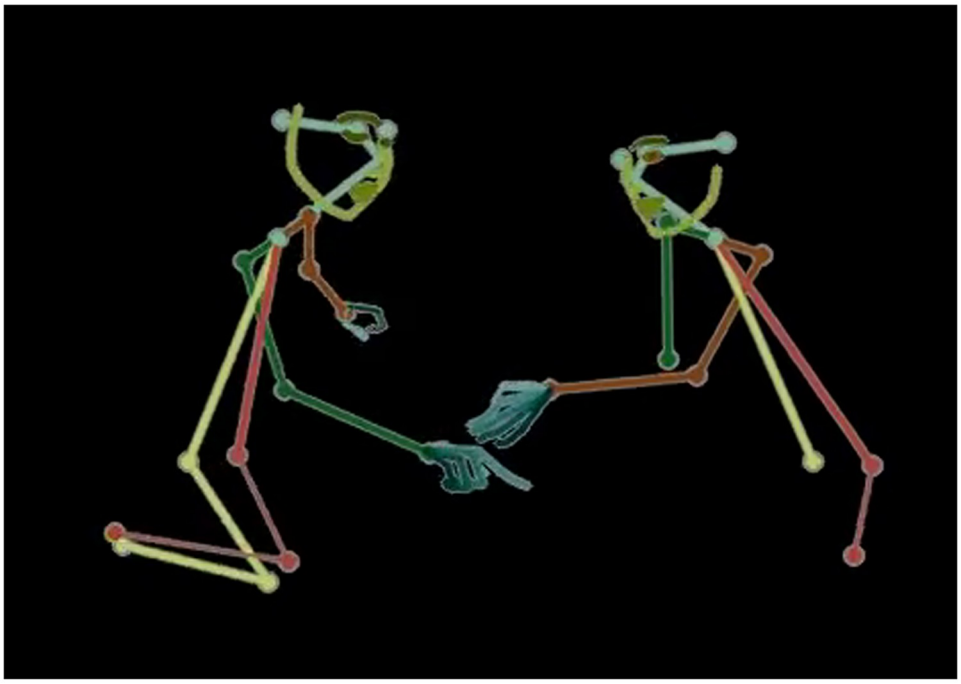


Fig 6. 2D skeletons, including facial landmarks and hand details are automatically extracted using the OpenPose library [18].

https://doi.org/10.1371/journal.pone.0205999.g006

providing a convenient interface to record the participants' demographics, resetting the game interface after each session, and automatically enforcing the acquisition protocol (presented in Table 1).

Coding of the social interactions

Our aim is to provide insights on the social dynamics, and as such we annotated the dataset using a combination of three coding schemes for social interactions that reuse and adapt established social scales. Our resulting coding scheme (Fig 5) looked specifically at three axis: the level of *task engagement* (that distinguishes between *focused*, *task oriented* behaviours, and *disengaged*—yet sometimes highly social – behaviours); the level of social engagement (reusing Parten's stages of play, but at a fine temporal granularity); the social attitude (that encoded attitudes like *supportive*, *aggressive*, *dominant*, *annoyed*, etc).

Task engagement. The first axis of our coding scheme aimed at making a broad distinction between 'on-task' behaviours (even though the free-play sandbox did not explicitly require the children to perform a specific task, they were still engaged in an underlying task: to play with the game) and 'off-task' behaviours. We called 'on-task' behaviours *goal oriented*: they encompassed considered, planned actions (that might be social or not). *Aimless* behaviours (with respect to the task) encompassed opposite behaviours: being silly, chatting about unrelated matters, having a good laugh, etc. These *Aimless* behaviours were in fact often highly social, and played an important role in establishing trust and cooperation between the peers. In that sense, we considered them as as important as on-task behaviours.

Social engagement: Parten's stages of play at micro-level. In our scheme, we characterised *Social engagement* by building upon Parten's stages of play [3]. These five stages of play

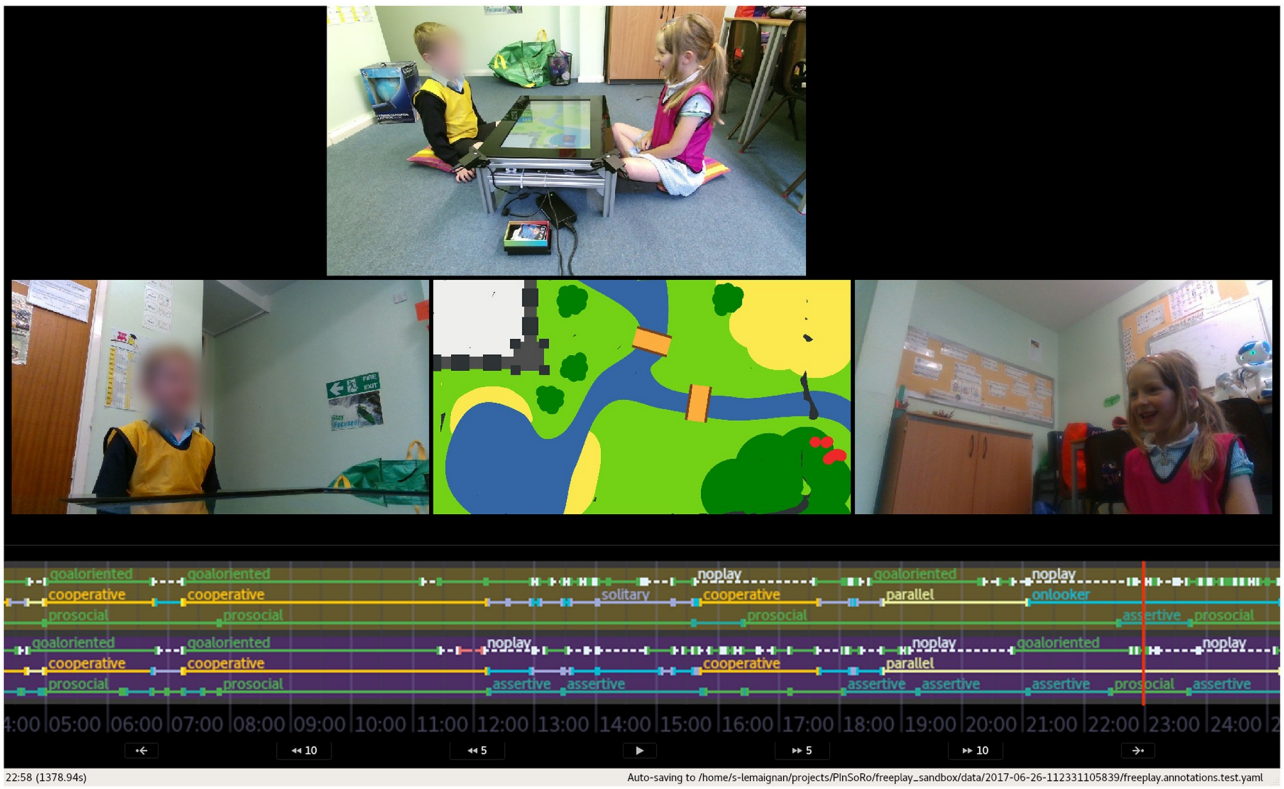


Fig 7. Screenshot of the dedicated tool developed for rapid annotation of the social interactions. The annotators used a secondary screen (tablet) with buttons (layout similar to Fig 5) to record the social constructs. Figure edited for legibility (timeline enlarged) and to mask out one of the children' face. The right individual pictured in this figure has given written informed consent (as outlined in PLOS consent form) to appear.

https://doi.org/10.1371/journal.pone.0205999.g007

are normally used to characterise rather long sequences (at least several minutes) of social interactions. In our coding scheme, we applied them at the level of each of the micro-sequences of the interactions: one child is drawing and the other is observing was labelled as *solitary play* for the former child, *on-looker* behaviour for the later; the two children discuss what to do next: this sequence was annotated as a *cooperative* behaviour; etc.

We chose this fine-grained coding of social engagement to enable proper analyses of the internal dynamics of a long sequence of social interaction.

Social attitude. The constructs related to the social *attitude* of the children derived from the *Social Communication Coding System* (SCCS) proposed by Olswang et al. [13]. The SCCS consists in 6 mutually exclusive constructs characterising social communication (*hostile*; *pro-social*; *assertive*; *passive*; *adult seeking*; *irrelevant*) and were specifically created to characterise children's communication in a classroom setting.

We transposed these constructs from the communication domain to the general behavioural domain, keeping the *pro-social*, *hostile* (whose scope we broadened in *adversarial*), *assertive* (i.e. dominant), and *passive* constructs. In our scheme, the *adult seeking* and *irrelevant* constructs belong to Task Engagement axis.

Finally, we added the construct *Frustrated* to describe children who are reluctant or refuse to engage in a specific phase of interaction because of a perceived lack of fairness or attention from their peer, or because they fail at achieving a particular task (like a drawing).

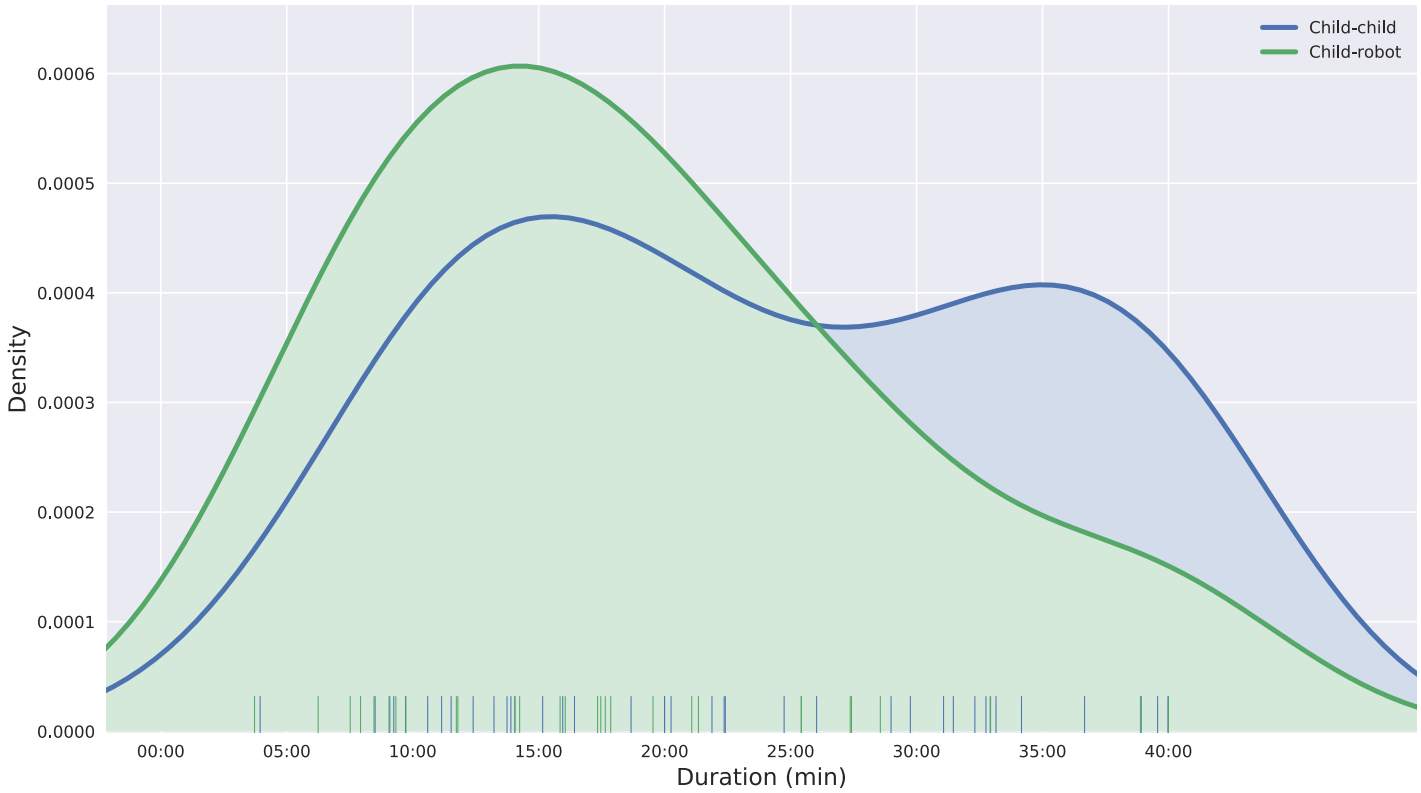


Fig 8. Density distribution of the durations of the interactions for the two conditions. Interactions in the child-robot condition were generally shorter than the child-child interactions. Interactions in the child-child condition followed a bi-modal distribution, with one mode centered around minute 15 (similar to the child-robot one) and one, much longer mode, at minute 37.

https://doi.org/10.1371/journal.pone.0205999.g008

Protocol

We adhered to the acquisition protocol described in Table 1 with all participants. To ease later identification, each child was also given a different and brightly coloured sports bib to wear.

Importantly, during the *Greetings* stage, we showed the robot both moving and speaking (for instance, "Hello, I'm Nao. Today I'll be playing with you. Exciting!" while waving at the children). This was of particular importance in the child-robot condition, as it set the children's expectations in term of the capabilities of the robot: the robot could in principle speak, move, and even behave in a social way.

Also, the game interface of the free-play sandbox offered a tutorial mode, used to ensure the children know how to manipulate items on a touchscreen and draw. In our experience, this never was an issue for children.

Data collection

Table 2 lists the raw datastreams that were collected during the game. By relying on ROS for the data acquisition (and in particular the `rosbag` tool), we ensured all the datastreams were synchronised, timestamped, and, where appropriate, came with calibration information (for the cameras mainly). For the PInSoRo dataset, cameras were configured to stream in qHD resolution (960×540 pixels) in an attempt to balance high enough resolution with tractable file size. It resulted in bag files weighting ≈1GB per minute.

Besides audio and video streams, user interactions with the game were monitored and recorded as well. The background drawings produced by the children were recorded. They

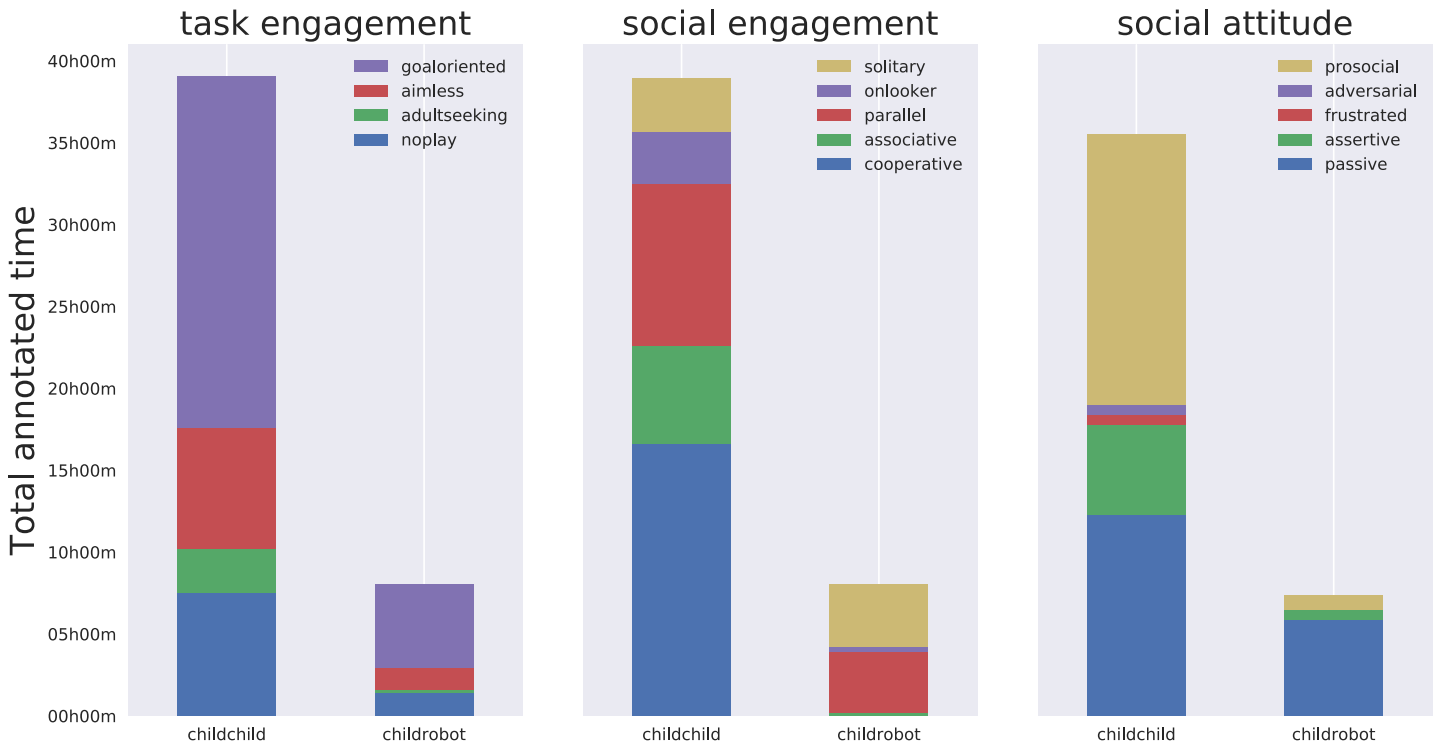


Fig 9. Repartition of annotations over the dataset (in total duration of recordings annotated with a given construct). The three classes of constructs (task engagement, social engagement, social attitude) and the two conditions (child-child and child-robot) are plotted separately.

https://doi.org/10.1371/journal.pone.0205999.g009

were also segmented according to their colours, and the contours of resulting regions were extracted and recorded. The positions of all manipulable game items were recorded (as ROS TF frames), as well as every touch on the touchscreen.

Data post-processing

Table 3 summarises the post-processed datastreams that are made available alongside the raw datastreams.

Audio processing. Audio features were automatically extracted using the OpenSMILE toolkit [14]. We used a 33ms-wide time windows in order to match the cameras FPS. We extracted the INTERSPEECH 2009 Emotion Challenge standardised features [15]. These are a range of prosodic, spectral and voice quality features that are arguably the most common features we might want to use for emotion recognition [16]. For a full list, please see [15]. As no reliable speech recognition engine for children voice could be found [17], audio recordings were not automatically transcribed.

Facial landmarks, action-units, skeletons, gaze. Offline post-processing was performed on the images obtained from the cameras. We relied on the CMU OpenPose library [18] to extract for each child the upper-body skeleton (18 points), 70 facial landmarks including the pupil position, as well as the hands' skeleton (Fig 6).

This skeletal information was extracted from the RGB streams of each of the three cameras, for every frame. It is stored alongside the main data in an easy-to-parse JSON file.

For each frame, 17 action units, with accompanying confidence levels, were also extracted using the OpenFace library [19]. The action-units recognised by OpenFace and provided

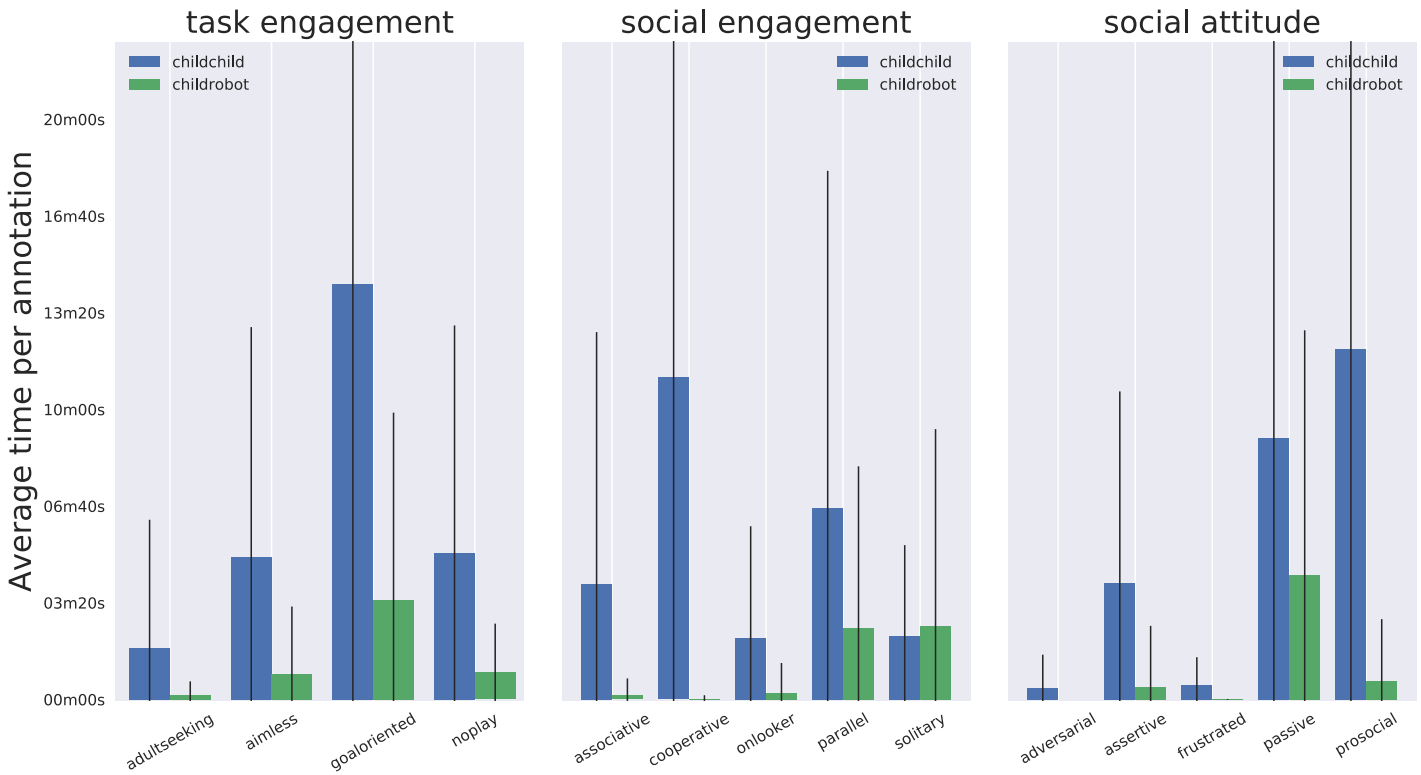


Fig 10. Mean time (and standard deviation) that each construct has been annotated in each recording. The large standard deviations reflect the broad range of group dynamics captured in the dataset.

https://doi.org/10.1371/journal.pone.0205999.g010

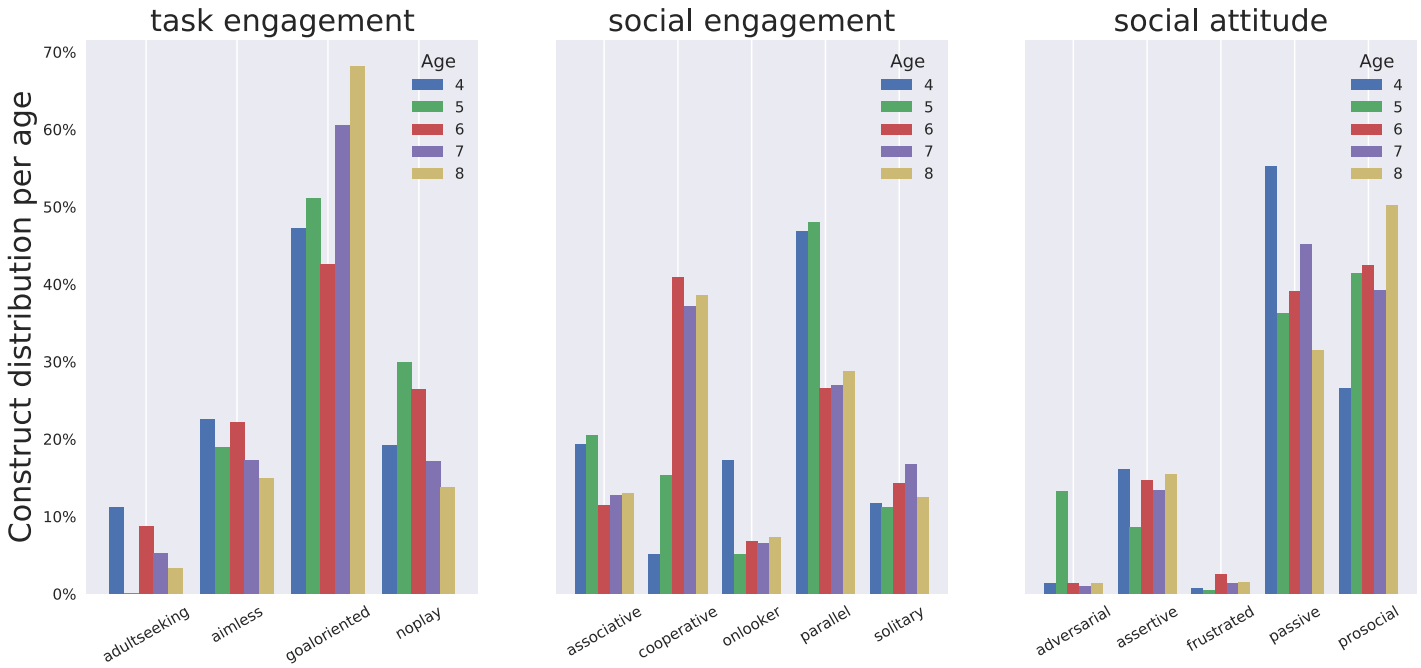


Fig 11. Percentage of observations for each constructs with respect the children's age.

https://doi.org/10.1371/journal.pone.0205999.g011

Table 1. Data acquisition protocol.

Greetings *(about 5 min)*
- explain the purpose of the study: showing robots how children play
- briefly present a Nao robot: the robot stands up, gives a short message (*Today I'll be watching you playing* in the child-child condition;*Today I'll be playing with you* in the child-robot condition), and sits down.
- place children on cushions
- complete demographics on the tablet
- remind the children that they can withdraw at anytime

Gaze tracking task *(40 sec)*
children are instructed to closely watch a small picture of a rocket that moves randomly on the screen. Recorded data is used to train a eye-tracker post-hoc.

Tutorial *(1-2 min)*
explain how to interact with the game, ensure the children are confident with the manipulation/drawing.

Free-play task (up to 40 min)
- initial prompt: *"Just to remind you, you can use the animals or draw. Whatever you like. If you run out of ideas, there's also an ideas box. For example, the first one is a zoo. You could draw a zoo or tell a story. When you get bored or don't want to play anymore, just let me know."*
- let children play
- once they wish to stop, stop recording

Debriefing *(about 2 min)*
- answer possible questions from the children
- give small reward (e.g. stickers) as a thank you

https://doi.org/10.1371/journal.pone.0205999.t001

Table 2. List of raw datastreams available in the PInSoRo dataset. Each datastream is timestamped with a synchronised clock to facilitate later analysis.

Domain	Type	Details
child 1	audio	16kHz, mono, semi-directional
	face (RGB)	qHD (960×540), 30Hz
	face (depth)	VGA (640×480), 30Hz
child 2	audio	16kHz, mono, semi-directional
	face (RGB)	qHD (960×540), 30Hz
	face (depth)	VGA (640×480), 30Hz
environment	RGB	qHD (960×540), 29.7Hz
game interactions	background drawing (RGB)	4Hz
	finger touches	6 points multi-touch, 10Hz
	game items pose	TF frames, 10Hz
other	static transforms between touchscreen and facial cameras	
	cameras calibration informations	

https://doi.org/10.1371/journal.pone.0205999.t002

alongside the data are AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28 and AU45 (classification following https://www.cs.cmu.edu/~face/facs.htm).

Gaze was also estimated, using two techniques. First, head pose estimation was performed following [20], and used to estimate gaze pose. While this technique is effective to segment pose at a coarse level (i.e. gaze on interactive table vs. gaze on other child/robot vs. gaze on experimenter), it offers limited accuracy when tracking the precise gaze location on the surface of the interactive table (due to not tracking the eye pupils).

We complemented head pose estimation with a neural network (a simple 7-layers, fully connected, multi-layer perceptron with ReLU activations and 64 units per layer), implemented

Table 3. List of post-processed datastreams available in the PInSoRo dataset. With the exception of social annotations, all the data was automatically computed from the raw datastreams at 30Hz.

Domain	Type	Details
children	face	70 facial landmarks (2D)
		17 facial action-units
		head pose estimation (TF frame)
		gaze estimation (TF frame)
	skeleton	18 points body pose (2D)
		20 points hand tracking (2D, only when visible)
	audio	INTERSPEECH's 16 low-level descriptors
annotations	timestamped annotations of social behaviours and remarkable events	

https://doi.org/10.1371/journal.pone.0205999.t003

with the Caffe framework (source available here: https://github.com/severin-lemaignan/visual_tracking_caffe).

The network trained from a ground truth mapping between the children' faces and 2D gaze coordinates. Training data is obtained by asking the children to follow a target on the screen for a short period of time before starting the main free play activity (see protocol, Table 1). The position of the target provides the ground truth (x, y) coordinates of the gaze on the screen. For each frame, the network is then fed a feature vector comprising 32 facial and skeletal (x, y) points of interest relevant to gaze estimation (namely, the 2D location of the pupils, eye contours, eyebrows, nose, neck, shoulders and ears). The training dataset comprises 80% of the fully randomized dataset (123711 frames) and the testing dataset the remaining 20% (30927 frames). Using this technique, we measured a gaze location error of 12.8% on our test data between the ground truth location of the target on the screen and the estimated gaze location (i.e. ±9cm over the 70cm-wide touchscreen). The same pre-trained network is then used to provide gaze estimation during the remainder of the free play activity.

Video coding. The coding was performed post-hoc with the help of a dedicated annotation tool (Fig 7) which is part of the free-play sandbox toolbox. This tool can replay and randomly seek in the three video streams, synchronised with the recorded state of the game (including the drawings as they were created). An interactive timeline displaying the annotations is also displayed.

The annotation tool offers a remote interface for the annotator (made of large buttons, and visually similar to Fig 5) that is typically displayed on a tablet and allow the simultaneous coding of the behaviours of the two children. Usual video coding practices (double-coding of a portion of the dataset and calculation of an inter-judge agreement score) were followed.

Results—The PInSoRo dataset

Using the free-play sandbox methodology, we have acquired a large dataset of social interactions between either pairs of children or one child and one robot. The data collection took place over a period of 3 months during Spring 2017.

In total, 120 children were recorded for a total duration of 45 hours and 48 minutes of data collection. These 120 children (see demographics in Table 4; sample drawn from local schools) were randomly assigned to one of two conditions: the child-child condition (90 children, 45 pairs) and a child-robot condition (30 children). The sample sizes were balanced in favour of the child-child condition as the social dynamics that we ultimately want to capture are much richer in this condition.

Table 4. Descriptive statistics for the children.

Condition	Age Mean	Age SD	# girls	# boys
Whole group	6.4	1.3	55	65
Child-child	6.3	1.4	42	48
Child-robot	6.9	0.9	12	18

https://doi.org/10.1371/journal.pone.0205999.t004

In both conditions, and after a short tutorial, the children were simply invited to freely play with the sandbox, for as long as they wished (with a cap at 40 min; cf. protocol in Table 1).

In the child-child condition, 45 free-play interactions (i.e. 90 children) were recorded with a mean duration M = 24.15 min (standard deviation SD = 11.25 min). In the child-robot condition, 30 children were recorded, M = 19.18 min (SD = 10 min).

Fig 8 presents the density distributions of the durations of the interactions for the two baselines. The distributions show that (1) the vast majority of children engaged easily and for non-trivial amounts of time with the task; (2) the task led to a wide range of levels of commitment, which is desirable: it supports the claim that the free-play sandbox is an effective paradigm to observe a range of different social behaviours; (3) many long interactions (>30 min) were observed, which is especially desirable to study social dynamics.

The distribution of the child-robot interaction durations shows that these interactions are generally shorter. This was expected as the robot's asocial behaviour was designed to be less engaging. Often, the child and the robot were found to be playing side-by-side—in some case for rather long periods of time—without interacting at all (solitary play).

Over the whole dataset, the children faces were detected on 98% of the images, which validates the positioning of the camera with respect to the children to record facial features.

Annotations

Five expert annotators performed the dataset annotation. Each annotator received one hour of training by the experimenters, and were compensated for their work.

In total, 13289 annotations of social dynamics were produced, resulting in an average of 149 annotations per record (SD = 136), which equates to an average of 4.2 annotations/min (SD = 2.1), and an average duration of annotated episodes of 48.8 sec (SD = 33.3). Fig 9 shows the repartition of the annotation corpus over the different constructs presented in Fig 5. Fig 10 shows the mean annotation time and standard deviation per recording for each construct.

Overall, 23% of the dataset was double-coded. Inter-coder agreement was found to be 51.8% (SD = 16.8) for task engagement annotations; 46.1% (SD = 24.2) for social engagement; 56.6% (SD = 22.9) for social attitude.

These values are relatively low (only partial agreement amongst coders). This was expected, as annotating social interactions beyond surface behaviours is indeed generally difficult. The observable, objective behaviours are typically the result of a superposition of the complex and non-observable underlying cognitive and emotional states. As such, these deeper socio-cognitive states can only be indirectly observed, and their labelling is typically error prone.

However, this is not anticipated to be a major issue for data-driven analyses, as machine learning algorithms are typically trained to estimate probability distributions. As such, divergences in human interpretations of a given social episode will simply be reflected in the probability distribution of the learnt model.

When looking at social behaviours with respect to age groups, expected behavioural trends are observed (Fig 11): *adult seeking* goes down when children get older; more *cooperative* play

is observed with older children, while more *parallel* play takes place with younger ones. In constrast, the social attitudes appear evenly distributed amongst age groups.

Dataset availability and data protection

All data has been collected by researchers at the University of Plymouth, under a protocol approved by the university ethics committee. The parents of the participants explicitly consented in writing to sharing of their child's video and audio with the research community. The data does not contain any identifying information, except the participant's images. The child's age and gender are also available. The parents of the children in this manuscript have given written informed consent (as outlined in PLOS consent form) to publish these case details.

The dataset is freely available to any interested researcher. Due to ethical and data protection regulations, the dataset is however made available in two forms: a public, Creative Commons licensed, version that does not include any video material of the children (no video nor audio streams), and hosted on the Zenodo open-data platform: https://zenodo.org/record/1043508. The complete version that includes all video streams is freely available as well, but interested researchers must first fill a data protection form. The detail of the procedure are available online: https://freeplay-sandbox.github.io/application.

Discussion of the free-play sandbox

The free-play sandbox elicits a loosely structured form of play: the actual play situations are not known beforehand and might change several times during the interaction; the game actions, even though based on one primary interaction modality (touches on the interactive table), are varied and unlimited (especially when considering the drawings); the social interactions between participants are multi-modal (speech, body postures, gestures, facial expressions, etc.) and unconstrained. This loose structure creates a fecund environment for children to express a range of complex, dynamics, natural social behaviours that are not tied to an overly constructed social situation. The diversity of the social behaviours that we have been able to capture can indeed been seen in Figs 9 and 11.

Yet, the interaction is nonetheless structured. First, the physical bounds of the interactive table limit the play area to a well defined and relatively small area. As a consequence, children are mostly static (they are sitting in front of the table) and their primary form of physical interaction is based on 2D manipulations on a screen.

Second, the game items themselves (visible in Fig 2) structure the game scenarios. They are iconic characters (animals or children) with strong semantics associated to them (such as 'crocodiles like water and eat children'). The game background, with its recognizable zones, also elicit a particular type of games (like building a zoo or pretending to explore the savannah).

These elements of structure (along with other, like the children demographics) arguably limit how general the PInSoRo dataset is. However, it also enable the free-play sandbox paradigm to retain key properties that makes it a practical and effective scientific tool: because the game builds on simple and universal play mechanics (drawings, pretend play with characters), the paradigm is essentially cross-cultural; because the sandbox is physically bounded and relatively small, it can be easily transported and practically deployed in a range of environments (schools, exhibitions, etc.); because the whole apparatus is well defined and relatively easy to duplicate (it essentially consists in one single touchscreen computer), the free-play sandbox facilitates the replication of studies while preserving ecological validity.

Compared to existing datasets of social interactions (the *Multimodal Dyadic Behavior Dataset*, the *Tower Game* dataset and the *UE-HRI* dataset), PInSoRo is much larger, with more than

45 hours of data, compared to 10.6, 5.6 and 6.9 hours respectively. PInSoRo is fully multimodal whereas the *Tower Game* dataset does not include verbal interactions, and the *UE-HRI* dataset focuses instead of spoken interactions. Compared to the *Multimodal Dyadic Behavior Dataset*, PInSoRo captures a broader range of social situations, with fully calibrated datastreams, enabling a broad range of automated data processing and machine learning applications. Finally, PInSoRo is also unique for being the first (open) dataset capturing *long sequences* (up to 40 minutes) of *ecologically valid* social interactions amongst children or between children and robots.

Conclusion—Towards the machine learning of social interactions?

We presented in this article the PInSoRo dataset, a large and open dataset of loosely constrained social interactions between children and robots. By relying on prolonged free-play episodes, we captured a rich set of naturally-occurring social interactions taking place between pairs of children or pairs of children and robots. We recorded an extensive set of calibrated and synchronised multimodal datastreams which can be used to mine and analyse the social behaviours of children. As such, this data provides a novel playground for the data-driven investigation and modelling of the social and developmental psychology of children.

The PInSoRo dataset also holds considerable promise for the automatic training of models of social behaviours, including implicit social dynamics (like rhythmic coupling, turn-taking), social attitudes, or engagement interpretation. As such, we foresee that the dataset might play an instrumental role in enabling artificial systems (and in particular, social robots) to recognise, interpret, and possibly, generate, socially congruent signals and behaviours whenever interacting with children. Whether such models can help uncover some of the implicit precursors of social behaviours, and is so, whether the same models, learnt from children data, can as well be used to interpret adult social behaviours, are open—and stimulating—questions that this dataset might contribute to answer.

Acknowledgments

The authors warmly thank the Plymouth's BabyLab, Freshlings nursery, Mount Street Primary School and Salisbury Road Primary School for their help with data acquisition. We also want to gratefully acknowledge the annotation work done by Lisa, Scott, Zoe, Rebecca and Sally.

This work has been supported by the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227) and the H2020 L2TOR project (grant 688014).

Author Contributions

Conceptualization: Séverin Lemaignan, Tony Belpaeme.

Data curation: Charlotte E. R. Edmunds.

Formal analysis: Charlotte E. R. Edmunds.

Funding acquisition: Tony Belpaeme.

Investigation: Séverin Lemaignan, Charlotte E. R. Edmunds.

Methodology: Séverin Lemaignan, Charlotte E. R. Edmunds.

Software: Séverin Lemaignan, Emmanuel Senft.

Supervision: Séverin Lemaignan, Tony Belpaeme.

Writing – original draft: Séverin Lemaignan.

Writing – review & editing: Séverin Lemaignan, Charlotte E. R. Edmunds, Emmanuel Senft.

References

1. Baxter P, Kennedy J, E S, Lemaignan S, Belpaeme T. From Characterising Three Years of HRI to Methodology and Reporting Recommendations. In: Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference (alt.HRI); 2016.
2. Bruner JS, Jolly A, Sylva K, editors. Play: Its role in development and evolution. Penguin; 1976.
3. Parten MB. Social participation among pre-school children. The Journal of Abnormal and Social Psychology. 1932; 27(3):243. https://doi.org/10.1037/h0074524
4. Nehaniv CL, Dautenhahn K. Imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions. Cambridge University Press; 2007.
5. Mohammad Y, Nishida T. Interaction Learning Through Imitation. In: Data Mining for Social Robotics. Springer; 2015. p. 255–273.
6. Nagai Y. Learning to comprehend deictic gestures in robots and human infants. In: Proc. of the 14th IEEE Int. Symp. on Robot and Human Interactive Communication. IEEE; 2005. p. 217–222.
7. Calinon S, Billard A. Teaching a humanoid robot to recognize and reproduce social cues. In: Proc. of the 15th IEEE Int. Symp. on Robot and Human Interactive Communication. IEEE; 2006. p. 346–351.
8. Liu P, Glas DF, Kanda T, Ishiguro H, Hagita N. How to Train Your Robot—Teaching service robots to reproduce human social behavior. In: Proc. of the 23rd IEEE Int. Symp. on Robot and Human Interactive Communication; 2014. p. 961–968.
9. Rehg J, Abowd G, Rozga A, Romero M, Clements M, Sclaroff S, et al. Decoding children's social behavior. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2013. p. 3414–3421.
10. Salter DA, Tamrakar A, Siddiquie B, Amer MR, Divakaran A, Lande B, et al. The tower game dataset: A multimodal dataset for analyzing social interaction predicates. In: Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on. IEEE; 2015. p. 656–662.
11. Ben-Youssef A, Clavel C, Essid S, Bilac M, Chamoux M, Lim A. UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM; 2017. p. 464–472.
12. Baxter P, Wood R, Belpaeme T. A touchscreen-based 'Sandtray'to facilitate, mediate and contextualise human-robot social interaction. In: Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on. IEEE; 2012. p. 105–106.
13. Olswang L, Svensson L, Coggins T, Beilinson J, Donaldson A. Reliability issues and solutions for coding social communication performance in classroom settings. Journal of Speech, Language & Hearing Research. 2006; 49(5):1058 – 1071. https://doi.org/10.1044/1092-4388(2006/075)
14. Eyben F, Weninger F, Gross F, Schuller B. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM international conference on Multimedia. May; 2013. p. 835–838. Available from: http://dl.acm.org/citation.cfm?doid=2502081.2502224.
15. Schuller B, Steidl S, Batliner A. The INTERSPEECH 2009 Emotion Challenge. In: Tenth Annual Conference of the International Speech Communication Association; 2009.
16. Schuller B, Batliner A, Seppi D, Steidl S, Vogt T, Wagner J, et al. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2007;2(101):881–884.
17. Kennedy J, Lemaignan S, Montassier C, Lavalade P, Irfan B, Papadopoulos F, et al. Child speech recognition in human-robot interaction: evaluations and recommendations. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. ACM; 2017. p. 82–90.
18. Cao Z, Simon T, Wei SE, Sheikh Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In: CVPR; 2017.
19. Baltrušaitis T, Mahmoud M, Robinson P. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. vol. 6. IEEE; 2015. p. 1–6.
20. Lemaignan S, Garcia F, Jacq A, Dillenbourg P. From Real-time Attention Assessment to "With-me-ness" in Human-Robot Interaction. In: Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference; 2016.

UNDERWORLDS: Cascading Situation Assessment for Robots

Séverin Lemaignan, Yoan Sallami, Christopher Wallbridge, Aurélie Clodic, Tony Belpaeme, Rachid Alami

UNDERWORLDS: Cascading Situation Assessment for Robots

Séverin Lemaignan[1], Yoan Sallami[2], Christopher Wallbridge[3],
Aurélie Clodic[2], Tony Belpaeme[3], and Rachid Alami[2]

***Abstract*— We introduce UNDERWORLDS, a novel lightweight framework for *cascading spatio-temporal situation assessment* in robotics. UNDERWORLDS allows programmers to represent the robot's environment as real-time distributed data structures, containing both scene graphs (for representation of 3D geometries) and timelines (for representation of temporal events). UNDERWORLDS supports *cascading* representations: the environment is viewed as a set of *worlds* that can each have different spatial and temporal granularities, and may inherit from each other. UNDERWORLDS also provides a set of high-level client libraries and tools to introspect and manipulate the environment models.**

This article presents the design and architecture of this open-source tool, and explores some applications, along with examples of use.

I. INTRODUCTION

UNDERWORLDS is a distributed and lightweight open-source framework[1] that enables robot programmers to build and refine spatial and temporal models of the environment surrounding a robot in real-time. UNDERWORLDS makes it possible to share these world models amongst the software components running on the robot. Additionally, UNDERWORLDS enables users to represent and manipulate *multiple alternatives* to the current, perceived world model in a distributed manner. For instance, the world with some objects filtered out; the world 'viewed' from the perspective of another agent; a hypothetical world resulting from the simulated application of a plan, etc.

A. Distributed Situation Assessment

Anchoring perceptions in a symbolic model suitable for decision-making requires perception abilities and their symbolic interpretation. We call *physical situation assessment* the cognitive skill that a robot exhibits when it represents and assesses the nature and content of its surroundings and monitors its evolution.

Numerous approaches exist, like amodal (in the sense of modality-independent) *proxies* [1], grounded amodal representations [2], semantic maps [3], [4], [5] or affordance-based planning and object classification [6], [7].

UNDERWORLDS is specifically inspired by geometric and temporal reasoners like SPARK (*SPAtial Reasoning & Knowledge*) [8] or TOASTER (*Tracking Of Agents and Spatio-TEmporal Reasoning*) [9]. SPARK acts as a situation assessment reasoner that generates symbolic knowledge from the geometry of the environment with respect to relations between objects, robots and humans. It also takes into account the different perspective that each agent has on the environment. SPARK embeds a modality-independent geometric model of the environment that serves both as basis for the fusion of the perception modalities and as bridge with the symbolic layer [10]. This geometric model is built from 3D CAD models of the objects, furniture and robots, and full body, rigged models of humans. It is updated at run-time by the robot's sensors. Likewise, UNDERWORLDS embeds a grounded amodal model of the environment, updated online from the robot's sensors (sensor fusion).

However, SPARK is a monolithic module that does not support sharing its internal 3D model with other external components. In contrast, UNDERWORLDS focuses on offering a shared and distributed representation of the environment within the robot's software architecture. This also distinguishes UNDERWORLDS from complex cognitive toolkits like KnowRob (as found in OpenEASE [11]). While these tools maintain a spatio-temporal model of the world, this model is internal and not meant to be made widely accessible to other external processes. UNDERWORLDS focuses instead on re-usability and sharing of distributed spatio-temporal models. As such, UNDERWORLDS can be seen as a middleware for spatio-temporal world models and, contrary to KnowRob, it does *not* provide any intrinsic high-level processing or reasoning capability. Such reasoning skills are implemented in loosely-coupled *clients* (see Section III hereafter).

Work on distributed scene graphs [12] has been previously applied to robotics to provide a shared 3D representation of the robot's environment (for instance, the *Robot Scene Graph* [13] or the *Deep State Representation* proposed in [14]). UNDERWORLDS offers a similar distribution mechanism for 3D scene graphs and extends it to temporal representations. Besides, UNDERWORLDS further extends this line of work by providing the ability to create, manipulate and share *multiple alternative worlds*. As an example, these could correspond to filtered or hypothetical *views* on the initial, perceived model of the environment.

B. Representing Alternative States of the World

The components which make use of spatial and temporal models of the environment are usually found in the intermediate layers of robotic architectures, between the low-level perceptual layers, and the high-level decisional layers. They include modules like geometric reasoners (that compute

[1]Author is with Bristol Robotics Lab, University of the West of England, Bristol, United Kingdom severin.lemaignan@brl.ac.uk, [2]Authors are with LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France firstname.surname@laas.fr, [3]Authors are with CRNS, Plymouth University, Plymouth, United Kingdom firstname.surname@plymouth.ac.uk

[1]https://github.com/underworlds-robot/underworlds

spatial and topological relations between objects), motion planners or action recognition modules.

These components exhibit different needs in terms of representation, like different nominal spatial and/or temporal resolutions. For instance, a 3D motion planner would typically use coarse 3D models of surrounding objects to lower the computational load while planning, while a module assessing the visibility of objects might need high-resolution models for accurate 3D visibility testing. This requirement of multiple task-specific representations has been framed as the need for *deep representations* by Beetz [15].

Traditional robotic middlewares, like ROS, are not particularly well suited to deal with these different needs: full geometric data can be represented, but is not first-class citizen: a basic task like displaying a 3D mesh at an arbitrary position is not particularly easy to perform with ROS, requiring the combination of static Collada meshes, a URDF kinematic description, TF broadcasters, and a 3D visualisation tool like RViz. Critically, simultaneously representing and reasoning on alternative states of the environment is not directly feasible.

Representing alternative states is however often highly desirable. For instance, software components manipulating environment models typically perform better if the models are physically consistent. However, low-level perception inaccuracies often introduce hard-to-avoid physical inconsistencies (like detected objects floating in the air, or wrongly inset into other objects). Therefore, a post-process stage (for instance, using a physics simulation engine) is needed to move the objects seen by the robot into physically-correct positions. Implemented with a classical approach (for instance, using ROS TF frames), we would represent an object `book` with two frames: the original frame (e.g.,`book_frame_raw`) and a second one computed by the physics engine (e.g.,`book_frame_corrected`). Such an approach leads to the robot's 3D model being cluttered with multiple frames and does not scale well.

Another example pertains to geometric task planning: a geometric task planner typically needs to reason over hypothetical future states of the environment ("What happens if I move this glass onto that pile of books?"). The planner generates many possible future states, which in turn might require further processing (for instance, running a physics simulation). Such a tool would benefit a flexible representation system, where models are derived from each other, with partial modifications and different timescales.

A third example relates to human-robot interaction scenarios where *perspective taking* is important (a prototypical example being the game 'I spy with my little eye', as implemented in [16]). Perspective taking is a cognitive skill that relies on the ability for an agent to take someone else's point of view to estimate what they see from their perspectives. Perspective taking has previously been implemented in robotics by temporarily placing virtual cameras at eye locations for each of the humans tracked by the robot [17]. While acceptable for simple cases, such an approach does not maintain truly independent spatio-temporal models of the environment for each agent, and in particular, it does not permit the representation of proper false-belief situations. On the contrary, separate, independent world models as implemented by UNDERWORLDS effectively support such a skill, which is an important precursor to research and implement human's mind modelling (i.e., a theory of mind) [18].

Lastly, geometric *pre-supposition accommodation* makes another interesting case for alternative worlds representation. *Pre-supposition accommodation* originally comes from linguistics, where it describes the mechanism by which *context is adjusted [...] to accept [...] a sentence that imposes certain requirements on the context in which it is processed* [19]. In the context of spatio-temporal representations, we call pre-supposition accommodation the ability of an agent to adjust its model so that it matches some contextual constraint. For instance, if A tells B to "catch the red balloon behind you", B might create a representation of an imaginary red balloon, placed behind her, even without actually observing the balloon: B *accommodates* the *pre-supposition* of a red balloon being present behind herself. Endowing robots with this capability has been touched upon by Mavridis et al. within their multi-modal *Grounded Situation Model* [2]. However, to the best of our knowledge, a general framework which would enable robots to accommodate spatial and temporal pre-suppositions by deriving imaginary worlds from existing ones has not been proposed so far.

UNDERWORLDS addresses this need and the main contribution of this work is a generic approach to **represent and share multiple parallel representations of the world**. UNDERWORLDS does so by allowing clients to clone existing worlds, modify them, and re-share them, without the cost of duplicating geometric data (as explained in section II). By organising clients in a network (Figure 1), worlds can be made dependent on each other, resulting in a loosely-coupled modular approach to spatio-temporal world representation that we call *cascading situation assessment*.

II. Design and Architecture

A. Software architecture

Figure 1 depicts a typical UNDERWORLDS topology: a graph (that happens to be an *acyclic* graph on Figure 1, but does not have to be in the general case) of worlds, with clients connecting the worlds to each others.

1) Clients: Software components implementing accessing UNDERWORLDS worlds are called clients. Clients can both read and write onto the worlds they are connected to, and automatically see updates broadcast by other clients connected to the same world. To ensure data consistency, worlds can have many simultaneous readers, but only one writer at a given time.

UNDERWORLDS provides several standard clients (like a 3D visualisation tool or a physics engine simulator). Clients are however typically written by the end users, depending on the needs of one's specific architecture.

2) Worlds: Worlds are effectively distributed data structures composed of a scene graph representing the 3D ge-

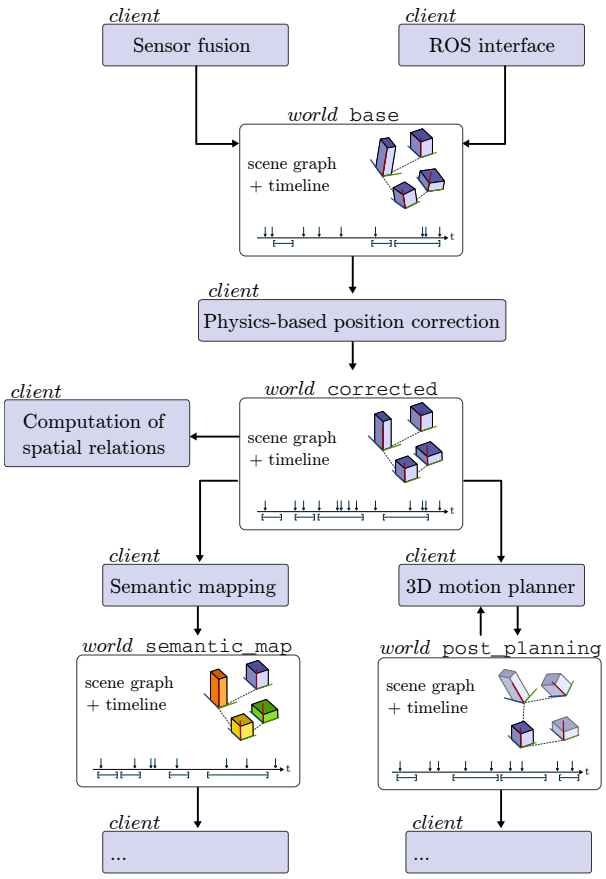


Fig. 1. Schema of a possible UNDERWORLDS network: eight *clients* (user-written & architecture specific; in blue) are sharing environment models through four independent *worlds* (made from joint spatial and temporal models). This architecture enables successive and modular refinement of the models (*cascading* situation assessment), effectively adapted to each client's needs.

ometry of the environment, and a timeline storing temporal events.

While each world is technically independent from all the others, dependencies (and therefore, coupling) arise between worlds from the clients' connections. For instance, filters effectively create a dependency between worlds. On Figure 1, the *Physics-based position correction* client creates a dependency between the world `base` (which represents here the result of raw sensor fusion) and the world `corrected` which would be a physically-consistent copy of `base`. As a result, an UNDERWORLDS network can also be seen as a dependency graph between worlds (where cyclic dependencies are permissible).

This architecture enables what we call *cascading situation assessment*: independent software components (the clients) build, refine and share successive models of the environment by a combination of filtering/transformations steps and model branching. A change performed by one client (for instance, a face tracker updates the pose of the human head) may thereby *cascade* to each of the downstream, dependent worlds.

3) Scenes: Worlds contain both a geometric model and a temporal model. The geometric model is represented as a scene graph. The scene graph has a unique root node, to which a tree of other nodes is parented.

Nodes in an UNDERWORLDS scene graph have three possible types: **objects** that represent concrete physical objects (typically with one or several associated 3D meshes); **entities** that represent abstract entities like reference frames or groups of objects; **perspectives** that represent viewpoints of the scene (like cameras or human gaze).

Every node has a unique ID, a parent, a 3D transformation relative to the parent and an optional name. *Object* nodes optionally store as well pointers to their associated meshes. Importantly, mesh data (or other geometric datasets like point clouds) are *not* stored within the nodes themselves. UNDERWORLDS represents geometric data as immutable data, identified by their hash value (preventing *de facto* data duplication). Nodes only store the hash corresponding to the desired geometric data, and the actual data is pulled from the server by the clients whenever they actually need it (for rendering for instance).

4) Timelines: Complementing the spatial representation encapsulated in the scene graph, each world also stores the world's *timeline*. This data structure is shared and synchronised amongst the clients in the same way as the scene graph. Clients can record and query both *events* (duration-less states) and *situations* in the timeline, i.e., states with a start time and a (possibly open-ended) end time.

B. Distributed spatio-temporal models

UNDERWORLDS is not a monolithic piece of software. Instead, it stands for both a *network of interconnected clients* which manipulate spatial and temporal models of the robot environment (for instance, a motion planner, a object detection module, a human skeleton tracker, etc.), and for a client library that makes it possible to interface existing software components with the network.

Critically, the network is essentially hidden to the client: from the user perspective, the environment model is manipulated as a local data structure (see Listing 1). Modifications to the model are asynchronously synchronised with a central server (the `underworlded` daemon) and broadcast to every other client connected to the same world.

As previously mentioned, worlds are composite data structures comprised of a scene graph and a timeline. These data structures are synchronised using Google's gRPC message passing framework[2], ensuring high throughput, reliability and cross-platform/cross-language support. The UNDERWORLDS API is specifically discussed hereafter, in section III-A.

UNDERWORLDS is meant to broadcast complex environment representations (typically including large geometric datasets, like meshes) in real-time. UNDERWORLDS itself does not perform many CPU intensive tasks (CPU intensive

[2] `http://www.grpc.io/`

processing tasks – sensor fusion, physics simulation, etc.– are performed by the clients themselves) and as such, the performance bottleneck is essentially the network's data throughput. In that regard, one of the simple yet critical optimisations performed by UNDERWORLDS is automatic caching of mesh data. Mesh data are not transmitted when nodes are updated; only a hash value of the mesh data. The client can then request the full data whenever it is actually needed.

C. Time and space complexity analysis

UNDERWORLDS is fundamentally about distributing two datastructures: a scene graph (with nodes representing spatial entities) and a timeline (where events are stored as a flat list). Typical time and space complexities arise from these datastructures. In typical usage scenarios (where the number of nodes or events remain under a few hundred relatively small), the computational load to manipulate these datastructures is however dominated by the actual processings performed by the clients with the data. In the current implementation, scene graphs and timelines are stored in-memory. Were they required, serialization and persistent storage are not anticipated to be difficult to implement.

More interesting is the time complexity of distributing changes across an UNDERWORLDS network. With n the number of worlds and m the number of clients in an UNDERWORLDS network, the worst-case (when every world is a parameter of every client) time complexity of creating or updating a node and propagating the change across the network is $O(n \times m)$ (this effectively corresponds to the UNDERWORLDS server performing $n \times m$ requests to notify clients of the update). The space complexity is the same (as clients own a full copy of the worlds they monitor), except for mesh data whose space and time complexities are $O(1)$ (only the server stores the mesh data).

In the common case of one client performing a full update of a single world (with p nodes) at each time step, the complexity of propagating these changes across the network would be $O(p \times m)$. Figure 2 shows measured propagation time for one change across up to 20 cascading worlds.

III. API & CLIENTS

A. API

As mentioned, UNDERWORLDS uses Google's gRPC as message passing protocol. The protocol is explicitly defined (using the *protocol buffers*[3] interface definition language), and bindings to various languages and platforms can be automatically generated from the protocol definition file (as of Jan 2018, gRPC can generate bindings for C, C++, C#, Node.js, PHP, Ruby, Python, Go and Java, on Windows, Mac, Linux and Android). The cross-platform/cross-language support of gRPC is especially welcome in the academic context, as it offers ease and flexibility to plug a variety of pre-existing components into an UNDERWORLDS network.

[3] `https://developers.google.com/protocol-buffers/`

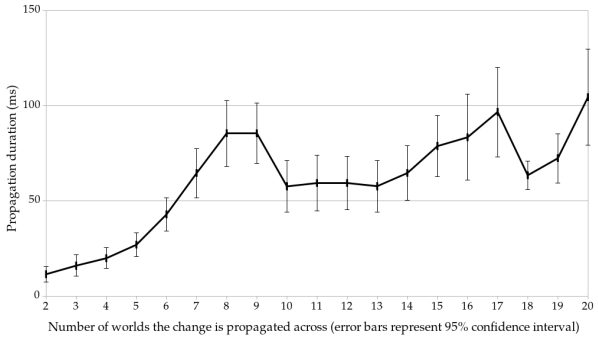


Fig. 2. Propagation times of one change (node creation) across n worlds. The test is performed by running $n-1$ pass-through filters that monitor one world and replicate any changes into the next world. Durations measured over 20 runs, performed on a 8 core machine.

However, the gRPC message passing layer is low-level with respect to the typical use of UNDERWORLDS (manipulation of asynchronous, distributed spatio-temporal models of the robot environment). In particular, the asynchronous fetching (and conversely, remote updating) of nodes and time-related objects is typically hidden from the user, and managed instead by the UNDERWORLDS client library.

UNDERWORLDS currently offers such a high-level client library for Python only (a C++ library is under development). Listing 1 gives a complete example of an UNDERWORLDS client performing simple filtering: the client continuously listens for changes in an input world, removes some objects (in this case, items whose volume is below a threshold), and forwards all other changes to an output world, effectively making the output world a copy of the input world with all smaller objects removed.

```
1  import underworlds
2
3  # by default, connect to the server on localhost
4  with underworlds.Context("small_object_filter") as ctx:
5
6      in_world = ctx.worlds["world1"]
7      out_world = ctx.worlds["world2"]
8
9      while True:
10
11         in_world.scene.waitforchanges()
12
13         for node in in_world.scene.nodes:
14             if node.volume > THRESHOLD:
15                 out_world.scene.nodes.update(node)
```

Listing 1: Example of a simple yet complete UNDERWORLDS filter, written in Python: the client connects to the UNDERWORLDS network, blocks until the world `world1` changes, and only propagate nodes that match the condition to the world `world2`.

B. Standard Clients

The UNDERWORLDS package provides several standard clients to perform common tasks on UNDERWORLDS networks.

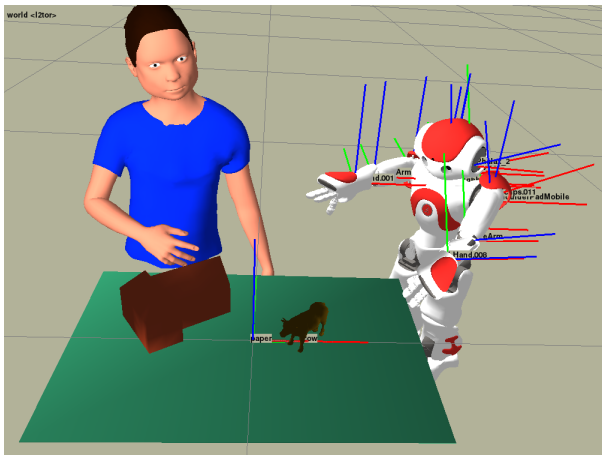


Fig. 3. Screenshot of the `uwds view` 3D visualisation and manipulation client. In this particular example, the 3D meshes have been pre-loaded using `uwds load`. Their positions are then updated at run-time using the robot's sensors and proprioception (joint state).

1) 3D Visualisation and manipulation: Interestingly, while UNDERWORLDS deals with 3D geometries and scenes, it does represent 3D entities purely as data structures; no visual representation is involved (and as such, the UNDERWORLDS server and core libraries do not depend on any graphics library like OpenGL). However, for all practical purposes, the ability to visualise the content of a scene is desirable. UNDERWORLDS provides a standard client, `uwds view`, that performs real-time 3D rendering of worlds, using OpenGL (Figure 3).

This tool also supports basic object manipulations (translations, rotations), that are broadcast to the other UNDERWORLDS clients connected to the same world.

Assets loading: Often, objects manipulated by the robot have known meshes with corresponding CAD models that can be conveniently pre-loaded. In these cases, UNDERWORLDS provides a tool, `uwds load`, that loads a mesh into a UNDERWORLDS network (and optionally, creates a node) from a large range of 3D formats (including Collada, FBX, OBJ, Blender)[4].

2) Physics simulation: When perception modules provide objects localisation, the physical consistency of the locations is not typically enforced. For instance, objects that are supposed to lay on a table might be slightly above (or inset into) the table; or when dropping an item into a box, the robot can not update the location of the item anymore as it becomes occluded.

These issues can be alleviated by relying on a physics simulation to stabilise the position of objects: natural physics (including gravity) are simulated for a short amount of time (up to one second) ahead of time, and the objects' positions are updated accordingly. To this end, UNDERWORLDS provide a standard filter, the `physics_filter`, based on the Bullet RT physics simulation and the `pybullet`[5] library. It generates an output world that mirrors its input world after a specific duration of physics simulation, the physical properties of objects (including mass, friction, inertia) being provided from standard URDF descriptions.

[4]The underlying import capability is provided by the ASSIMP library. `http://assimp.sourceforge.net/`

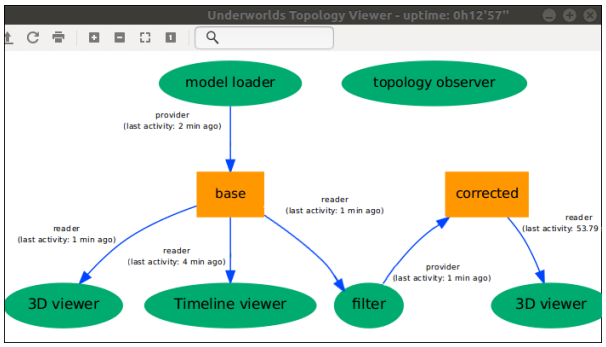


Fig. 4. Screenshot of the network topology introspection tool, with arbitrary examples of worlds (represented as boxes) and clients (ellipses). CLients are connected to the worlds either as *readers* or *providers* of data. UNDERWORLDS introspection features make it possible to also visualise when clients were last active.

3) Introspection and debugging: UNDERWORLDS provides a range of tools to inspect a running network. Graphical tools (`uwds explorer` and `uwds timeline`, see Figure 4) provide a user-friendly overview of the system's graph with the connections between the clients and the worlds, as well as their activity.

Specialised command-line tools are also available to list the worlds and their content (`uwds ls`) at run-time, or to display detailed information for a specific node (`uwds show`).

4) Interface with ROS: UNDERWORLDS is meant to integrate as easily as possible into existing robot architectures, and interfaces transparently with ROS' TF frame system through the `uwds tf` client.

The `uwds tf` client continuously monitors the ROS TF tree, and mirrors TF frames as nodes in the desired UNDERWORLDS world. A node is first created if none matches a given TF frame, and its transformation is subsequently updated, mirroring the TF frame. A regular expression can be provided to only mirror a subset of the TF tree into UNDERWORLDS.

Currently, the process is unidirectional: the `uwds tf` client performs TF to UNDERWORLDS updates, but not the reverse.

C. Spatial Reasoning and Perspective Taking

Spatial reasoning [20] is a field in its own right, and has been used for natural language processing for applications such as direction recognition [21], [22] or language grounding [23]. Other examples in human-robot interaction include

[5]`https://pybullet.org/`

Ros et al. [17], [16] which has recently been integrated into a full architecture for autonomous human-robot interaction [10].

UNDERWORLDS provides an exemplary client (`spatial_relations`) to compute both allo-centric (independent of the viewpoint like `isIn` or `isOn`) and ego-centric (i.e., viewer-dependent, like `inFrontOf` or `leftOf`) spatial relations between objects. Other libraries, like QSRLib [24], that implement computational models of Qualitative Spatial Relations, could be trivially combined with UNDERWORLDS to provide more advanced geometric analysis. Future developments will also include the results of the more basic research on spatio-temporal reasoning for robotics, led by de Leng and Heintz [25].

UNDERWORLDS also implements an efficient algorithm to assess object visibility from a specific viewpoint (i.e., from a given *perspective* node). The algorithm (color picking) enables fast (single pass) computation of the visibility of every object in the scene, while providing control regarding how many pixels should be actually visible for the object to be considered globally visible. The command-line tool `uwds visibility` returns the list of visible objects from the point of view of each camera in a given world, and UNDERWORLDS also provides the helper class `VisibilityMonitor` to programmatically access visibility information.

When integrated into a filter node, visibility computation allows easy creation of new worlds representing the estimated perspectives of the different agents.

IV. Application Example: Perspective-aware joint actions

UNDERWORLDS is being used within the large European project MuMMER[6] for service robots to compute visibility and knowledge about objects, places and agents within a mall environment.

We present here a simplified scenario, yet representative of situations which are processed in real-time by MuMMER robots: two humans and a robot are looking at a table and have to coordinate joint actions (pick and place). One object on the table (the green box in Figure 5) is only visible to one human and the robot, but hidden to the second human. The robot needs to take into account this fact to generate appropriate and legible joint manipulation actions. Figure 5 illustrates the topology of the UNDERWORLDS network that we use to this end.

A first client, *static_env_provider*, provides the environment models and allows to build a first `ENV` world where static objects, furnitures and walls are present. Then, three worlds cascade through three (independent) clients: *robots_state_monitor* augments `ENV` with the robot state (using underneath the ROS robot state publisher node) and broadcast a new world `ENV_ROBOTS`. *objects_monitor* then recognises and adds the dynamic objects (using `ar_track_alvar`[7]). *humans_monitor* finally detects and

[6] `http://www.mummer-project.eu`
[7] `http://wiki.ros.org/ar_track_alvar`

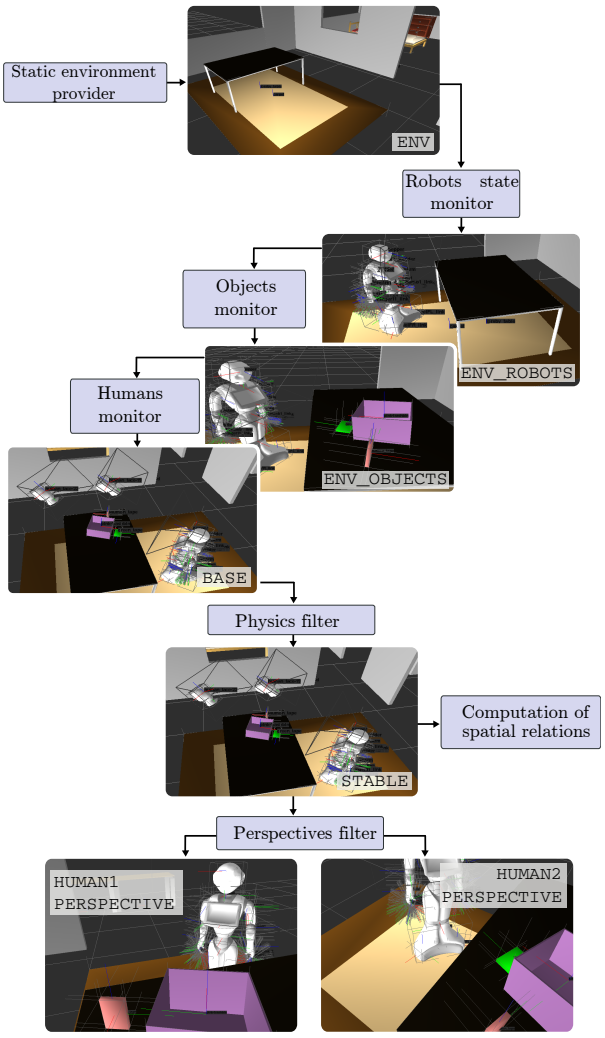


Fig. 5. Schema of the UNDERWORLDS architecture used in the MuMMER project. Clients read and generate the worlds `ENV` $\rightarrow$ `ENV_ROBOT` $\rightarrow \ldots \rightarrow$ `HUMAN*_PERSPECTIVE`. The last two worlds `HUMAN{1,2}_PERSPECTIVE` represent the immediate visual perspective of each of the humans, as well as their past visual perceptions. As such, they are the *visual memories* of the humans, that the robot can rely on when making decisions.

continuously updates the humans poses (using [26]). It broadcasts a world called `BASE` that contains as a result the static environment, the robots, the dynamic objects and the detected humans.

The world `BASE` goes through a *physics filter* client (as explained in section III-B.2) to obtain the `STABLE` world where all elements are present with physically-consistent locations. This physically-correct world is used by the *computation_of_spatial_relations* client to compute spatial relations such as `onTop`, `isIn` or `isAbove` (see Section III-C).

The world `STABLE` is also used by a *perspectives_filter* client to compute the different visual perspectives of each agent (in our case: human 1, human 2 and the robot itself).

In addition to a 3D rendering of the input world from the perspective of the agent, it aggregates the history of what was visible to the agent at a given point in time. As such, it does not only offer a snapshot of the agent visual perspective at the current time but also acts as the visual memory of each agent.

With this network, the robot can easily compute that an object on the table is only seen by the human 1 and not the human 2; additionally, if human 1 moves in a position where the object is not visible anymore to him, the *perspective_filter* will maintain the knowledge that the human had seen it (and keep the last position where it has been seen).

UNDERWORLDS makes it possible to implement such a geometric reasoning pipeline in a fully decoupled way, and each intermediary world can be easily introspected at run-time. This example shows how UNDERWORLDS facilitates the implementation and debugging of complex spatio-temporal reasoning pipelines.

We are currently deploying a similar network in the framework of the European project MuMMER where a Pepper robot handles interactive situations in a large shopping centre in Finland. One of the situation is a guiding task where Pepper help people to find their route by pointing them landmarks and explaining them how to reach a destination. To be effective, this helping behaviour needs to be aware of the visual perspective of the human. UNDERWORLDS facilitates the implementation of such a spatio-temporal reasoning pipeline, where perception and high-level reasoning (including complex, human-aware reasoning) have to be tightly integrated. Because of the decoupling of each of the clients in the network, UNDERWORLDS also practically supports software development spread across multiple partners in different countries, with different expertise.

V. Discussion and Conclusion

A. Relation to existing robotic middleware

Like traditional robotic middleware, UNDERWORLDS offers a form of distributed computation based on message passing. However, it distinguishes itself from existing middlewares (including ROS extensions like DyKnow [27]) in significant ways. Most importantly, UNDERWORLDS purposefully does not offer any *general* capability to distribute computation and data streams amongst independent components: it focusses specifically on distributing environment models, both spatial (geometric models) and temporal (events and situations). In that sense, UNDERWORLDS really is a *distributed datastructure* that addresses the specific needs of spatio-temporal modelling, including the modelling of hypothetical, alternative world models, something that traditional middlewares like ROS do not address adequately. Second, and as presented above, UNDERWORLDS offers specific mechanisms for the representation and manipulation of alternative world models that are not directly achievable with traditional tools.

While using standard middleware as *underlying transport* for UNDERWORLDS would be technically feasible and relatively easy to implement, it does not offer any clear advantage over lighter and dedicated message passing libraries like ZeroMQ or gRPC (the later being the one used by UNDERWORLDS).

B. Future work

As illustrated in section IV, UNDERWORLDS is already deployed and used on the field. Several features are however still under development.

1) Representation capabilities: as presented in section II, the current version of UNDERWORLDS allows to represent *objects*, *abstract entities* like groups and *perspectives*. *Fields* are also part of the UNDERWORLDS design, but are not yet implemented. Fields are commonly used to represent continuously-valued spatial entities. Fields might or might not be spatially bounded. Examples include the working space of a robot arm (spatially bounded), the field of view of a camera (spatially bounded), proxemics (potentially unbounded). We plan to represent fields in UNDERWORLDS using the memory-efficient octomaps [28] or NDT-OM maps [29]. Similarly to geometric data,these datastructures will not be directly stored with the nodes (nodes will refer to them through handles), but unlike geometric data, they will not be treated as immutable datasets by the server, permitting real-time updates.

Representation of uncertainty: currently node positions are stored as 4×4 transformation matrices, relative to the node parent. This representation is efficient, and conveniently matches traditional representation systems (including ROS TF frames or OpenGL transformations). However, the explicit management of uncertainties is instrumental to many robotic applications, and we plan to add full support for position uncertainties to UNDERWORLDS. We plan to add this support by adding a pose covariance matrix to the nodes, and equipping the different UNDERWORLDS helper tools with corresponding support (like covariance ellipses visualisation in `uwds view`).

2) Implementation and Integration: we plan to continue to improve the integration of UNDERWORLDS into existing software architectures. A short-term goal is to provide excellent C++ support, with a high-level, user-friendly C++ client library. This is critical for a broader adoption of UNDERWORLDS within the robot community. Support for other languages might follow, depending on demands and open-source contributions.

C. Conclusion

We have introduced UNDERWORLDS, a novel framework for shared and composable spatio-temporal representations of a robot's world. The key contributions of our approach are: a composite data structure for environment representation within a robotic software architecture, made of a scene graph and a timeline; a mechanism to efficiently and transparently share this data structure amongst a set of clients (the software modules of the robot); a cascading architecture permitting the explicit of representation of alternative states of the world while maintaining a network of dependencies.

We have additionally presented a concrete instantiation of a system relying on UNDERWORLDS for its representation needs, and we have sketched future directions of development.

We believe this work can practically support existing robotic architectures with state-of-the-art spatio-temporal representation capabilities. We also hope that this line of research can lead to a better understanding of the representation needs of modern robotic systems, and participate to the emergence of a possible common representation platform for robots, building on previous formalisation efforts like the RSG-DSL domain specific language [30].

ACKNOWLEDGMENT

This work has been supported by the EU H2020 Marie Skłodowska-Curie Actions project DoRoThy (grant 657227), the EU H2020 MuMMER project (grant 688147) and the EU H2020 L2TOR project (grant 688014).

REFERENCES

[1] H. Jacobsson, N. Hawes, G.-J. Kruijff, and J. Wyatt, "Crossmodal content binding in information-processing architectures," in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*. New York, NY, USA: ACM, 2008, pp. 81–88.

[2] N. Mavridis and D. Roy, "Grounded situation models for robots: Where words and percepts meet," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006.

[3] A. Nüchter and J. Hertzberg, "Towards semantic maps for mobile robots," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 915 – 926, 2008.

[4] C. Galindo, J. Fernández-Madrigal, J. González, and A. Saffiotti, "Robot task planning using semantic maps," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 955–966, 2008.

[5] N. Blodow, L. C. Goron, Z.-C. Marton, D. Pangercic, T. Rühr, M. Tenorth, and M. Beetz, "Autonomous semantic mapping for robots performing everyday manipulation tasks in kitchen environments," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Francisco, CA, USA, September, 25–30 2011.

[6] C. Lörken and J. Hertzberg, "Grounding planning operators by affordances," in *International Conference on Cognitive Systems (CogSys)*, 2008, pp. 79–84.

[7] K. Varadarajan and M. Vincze, "Ontological knowledge management framework for grasping and manipulation," in *IROS Workshop: Knowledge Representation for Autonomous Robots*, 2011.

[8] E. A. Sisbot, R. Ros, and R. Alami, "Situation assessment for human-robot interactive object manipulation," in *2011 RO-MAN*, July 2011, pp. 15–20.

[9] G. Milliez, M. Warnier, A. Clodic, and R. Alami, "A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 1103–1109.

[10] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami, "Artificial cognition for social human-robot interaction: An implementation," *Artificial Intelligence*, 2016.

[11] M. Beetz, M. Tenorth, and J. Winkler, "Open-EASE – a knowledge processing service for robots and robotics/ai researchers," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1983–1990.

[12] M. Naef, E. Lamboray, O. Staadt, and M. Gross, "The blue-c distributed scene graph," in *Proceedings of the workshop on Virtual environments 2003*. ACM, 2003, pp. 125–133.

[13] S. Blumenthal, H. Bruyninckx, W. Nowak, and E. Prassler, "A scene graph based shared 3d world model for robotic applications," in *2013 IEEE International Conference on Robotics and Automation*, May 2013, pp. 453–460.

[14] P. Bustos, L. J. Manso, J. P. Bandera, A. Romero-Garcés, L. V. Calderita, R. Marfil, and A. Bandera, "A unified internal representation of the outer world for social robotics," in *Robot 2015: Second Iberian Robotics Conference*. Springer, 2016, pp. 733–744.

[15] M. Beetz, D. Jain, L. Mösenlechner, and M. Tenorth, "Towards performing everyday manipulation activities," *Robotics and Autonomous Systems*, vol. 58, no. 9, pp. 1085–1095, 2010.

[16] R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Which one? grounding the referent based on efficient human-robot interaction," in *19th IEEE International Symposium in Robot and Human Interactive Communication*, 2010.

[17] R. Ros, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Solving ambiguities with perspective taking," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2010, pp. 181–182.

[18] S. Lemaignan and P. Dillenbourg, "Mutual modelling in robotics: Inspirations for the next steps," in *Proceedings of the 2015 ACM/IEEE Human-Robot Interaction Conference*, 2015.

[19] K. Von Fintel, "What is presupposition accommodation, again?" *Philosophical perspectives*, vol. 22, no. 1, pp. 137–170, 2008.

[20] J. O'Keefe, *The Spatial Prepositions*. MIT Press, 1999.

[21] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *HRI*, 2010, pp. 259–266.

[22] C. Matuszek, D. Fox, and K. Koscher, "Following directions using statistical machine translation," in *Proceedings of the International Conference on Human-Robot Interaction*. ACM Press, 2010.

[23] S. Tellex, "Natural language and spatial reasoning," Ph.D. dissertation, MIT, 2010.

[24] Y. Gatsoulis, M. Alomari, C. Burbridge, C. Dondrup, P. Duckworth, P. Lightbody, M. Hanheide, N. Hawes, D. Hogg, A. Cohn *et al.*, "Qsrlib: a software library for online acquisition of qualitative spatial relations from video," Tech. Rep., 2016.

[25] D. de Leng and F. Heintz, "Qualitative spatio-temporal stream reasoning with unobservable intertemporal spatial relations using landmarks." in *AAAI*, 2016, pp. 957–963.

[26] V. Khalidov and J.-M. Odobez, "Real-time multiple head tracking using texture and colour cues," Idiap, Idiap-RR Idiap-RR-02-2017, 2 2017.

[27] D. de Leng and F. Heintz, "DyKnow: A Dynamically Reconfigurable Stream Reasoning Framework as an Extension to the Robot Operating System," in *IEEE Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR)*, 2016, pp. 55–60. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-132266

[28] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: An efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, 2013.

[29] J. P. Saarinen, H. Andreasson, T. Stoyanov, and A. J. Lilienthal, "3d normal distributions transform occupancy maps: An efficient representation for mapping in dynamic environments," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1627–1644, 2013.

[30] S. Blumenthal and H. Bruyninckx, "Towards a domain specific language for a scene graph based robotic world model," *arXiv preprint arXiv:1408.0200*, 2014.

This is the final draft post-refereeing of Vollmer, A. L., Read, R., Trippas, D., & Belpaeme, T. (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. Science Robotics, 3(21), eaat7111.

For a copy of the pdf of the published paper please visit http://robotics.sciencemag.org/content/3/21/eaat7111, or email a request to avollmer@techfak.uni-bielefeld.de or tony.belpaeme@ugent.be.

Children conform, adults resist: robot group induced peer pressure on normative social conformity

Anna-Lisa Vollmer,[1*] Robin Read,[2] Dries Trippas,[3] Tony Belpaeme[2,4*]

[1]Bielefeld University, Cluster of Excellence Cognitive Interaction Technology, 33619 Bielefeld, Germany
[2]Plymouth University, Centre for Robotics and Neural Systems, Plymouth, PL4 8AA, United Kingdom
[3]Max Planck Institute for Human Development, Center for Adaptive Rationality, 14195 Berlin, Germany
[4]Ghent University, IDLab – imec, B-9052 Ghent, Belgium

[*]To whom correspondence should be addressed; E-mail: avollmer@techfak.uni-bielefeld.de.

People are known to change their behavior and decisions in order to conform to others, even for obviously incorrect facts. Due to recent developments in artificial intelligence and robotics, robots increasingly are found in human environments and there they form a novel social presence. It is as yet unclear if and to what extent these social robots are able to exert similar peer pressure. This study uses the Asch paradigm which shows how participants conform to others while performing a visual judgment task. We first replicate the finding that adults are influenced by their peers, but show that they resist social pressure from a group of small humanoid robots. Next, we repeat the study with 7 to 9-year old children and show that children do conform to the robots. This raises opportunities as well as concerns for the use of social robots with young and vulnerable cross-sections of society; while conforming can be beneficial, the potential for misuse and the potential impact of erroneous performance

cannot be ignored.

One-sentence summary

Children show increased yielding to social pressure exerted by a group of robots, adults however resist being influenced by our robots.

Introduction

Social robots represent a new frontier in the personal robotics industry. These robots are designed to autonomously interact with people across a variety of different application domains in natural and intuitive ways, using the same repertoire of social signals used by humans (*1–3*). Current applications include robotic tour guides in museums (*4*), therapeutic aids in care homes (*5*) and early years childcare (*6, 7*), and teaching aids in primary school classrooms (*2, 8, 9*), with future applications forecast to be far broader (*10*). With these future applications, robots will share the same physical and social space as users, which raises questions regarding safety, and given the social nature of the robots, the psychosocial impact.

It has been shown that people, particularly the younger age groups, easily form strong bonds with social robots, so much so that it can cause distress when a robot is mistreated or misbehaves (*6, 11*), even when they are crude approximations to real living organisms (*12*). Conversely, interaction with social robots has also been found to elicit and reinforce healthy social behaviors in children with autism spectrum disorder (*13–15*) as well as promote and augment social behavior and bonding between group members in care homes (*5*). An open question is whether these social bonds offer robots other affordances such as the ability to exert social influence (*16*), and whether people yield to these.

The *computers as social actors (CASA)* hypothesis (*17–19*) states that people naturally and unconsciously treat computers and other forms of media in a manner that is fundamentally

social, attributing human-like qualities to technology. It has had a notable impact in the fields of Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI). Assuming that the CASA hypothesis holds true, it predicts that people, regardless of their age, are sensitive to (and submit to) social influences exerted by social robots and (crucially) that this is automatic and involuntary (*18*). We tested this prediction by replicating the influential paradigm to study normative social conformity devised by Solomon Asch (*20–22*).

Computers as social actors

Reeves and Nass concluded from a number of social psychology experiments that "individuals' interactions with computers, television, and new media are fundamentally social and natural, just like interactions in real life." (*17, p. 5*). The CASA hypothesis is part of the Media Equation hypothesis (*17*), an overarching theory which additionally implies that people process experiences mediated by technology in the same way as they process unmediated experiences. Describing an unconscious and automatic response, the CASA hypothesis seems to apply to everyone regardless of expertise.

The studies conducted by Reeves and Nass show that people treat technology like people, using the same social rules, expectations, beliefs and behaviors towards technology as they would with other people, according them social behaviors (e.g., politeness, reciprocity), attributing human characteristics to them (e.g., gender), reacting to them as they would to human interaction partners, and so on (*18, 19*). Nass and colleagues found that when a computer asks a user to evaluate itself, the user will give more positive feedback than when the user does the evaluation on a different computer (*23*). They also found that people showed gender stereotypes toward computers with male and female voice (*24*). Rules of attraction seem to hold as well. Users were shown to like electronic partners better when they have the same personality as the user (*17*).

Peer-driven normative conformity and the Asch paradigm

Conformity describes the behavior of an individual who is complying with group norms. In the field of social psychology, two main varieties of conformity are considered: informational social conformity and normative social conformity. The former depicts the influence of others' responses as a source of information on one's own judgment when a task is ambiguous and the correct answer not straightforward. The latter describes an influence of others on judgments in a task with unambiguous stimuli where the correct answers are clear. Participants are lead to give incorrect responses complying publicly with an erroneous majority in order to be accepted.

The well-established and most influential paradigm to study normative social influence was devised by Solomon Asch in 1951 (*20*). In his classic conformity experiments, individual participants were unknowingly grouped with multiple confederates and instructed to judge the length of a target line compared to three comparison lines, only one of which has the same length as the target line (Fig. 1D). For each such comparison, all the participants verbally reported one after the other which comparison line they perceived to match the target line, with the subject verbalizing their answer before the last of the confederates. On two-thirds of the trials the confederates unanimously announced an incorrect judgment (critical trials, $n = 12$) while providing the correct response on the remaining trials (neutral trials, $n = 6$). The participants followed the group response, complying publicly and submitting to group pressure in 32% of trials (in 68% of critical trials they responded correctly; one fourth of the participants were completely independent and resisted the group pressure in all critical trials) (*20*). Asch conducted his first experiment with male college students and a majority group of varying size.

Many replications and alterations of this standard experiment have been conducted to identify factors that influence conformity. Size, immediacy, unanimity, and personal importance of the group, the ambiguity and public announcement of responses, gender, and age are among these factors. Whereas conformity seems to increase with a larger majority, it changes only lit-

tle from group sizes of four (*20, 22*). Majority groups that are personally more important to the participant (e.g. peers, in-group vs. out-group members) (*25, 26*) exert a greater social pressure. If there is only one dissenter in the majority group who announces the correct or even only a different answer from the group, conformity decreases drastically (*21*). It increases as the correct judgment becomes more ambiguous (e.g., by making the line lengths more similar) (*22*). Participants that write down their judgments privately tend to resist group pressure (*22*). Female participants were found to endorse the group response slightly more often than male participants (*27, 28*). Age has been reported to reduce susceptibility to social influence (*29, 30*), although findings seem to be conflicting (*31, 32*).

Results

We have tested whether adults (Experiment 1) and children (Experiment 2) exhibit normative social conformity (*16*) when conducting a visual discrimination task in the presence of three humanoid robots (Fig. 1, A–C). We replicated the Asch paradigm to study normative social conformity. The original group setup formed the basis of our experimental condition. As a control condition, participants were asked to perform the same task while alone. Decreased accuracy on the critical trials in the experimental condition compared to the control condition is evidence for social conformity.

FIGURE 1 ABOUT HERE

Adults

In Experiment 1 we tested the hypothesis that humanoid robots exert normative social pressure on adults. Participants ($N = 60$, 34 female, age: range = 18 – 69 years, $M = 30.9$ years, $SD = 14.2$) were randomly assigned to one of three conditions: a control condition ($n = 20$), a 'human peer' condition ($n = 20$) with three human confederates, and a 'robot peer' condition

($n = 20$) in which three humanoid robots replaced the human confederates.

In all conditions, participants, including the confederates in the human-peer condition and robots in the robot-peer condition, were asked to verbally report which line matched the reference line. The experimenter decided on the response order.

On each trial we measured whether the real participant's verbal response was correct. The experiment was a 3 (condition: control vs. human peer vs. robot peer, between subjects) $\times$ 2 (trial type: critical vs. neutral, within subjects) mixed design. If people are influenced by social peers, line judgment accuracy in the critical (but not the neutral) trials should be lower for the peer conditions compared to the control condition.

Analysis of Logistic Regression model

There was a significant main effect of condition ($\chi^2(2) = 11.8$, $P = .003$), suggesting that peers influenced line judgment accuracy. The condition main effect was qualified by an interaction with trial type, $\chi^2(2) = 11.9$, $P = .003$, indicating that the effect of peers differed for the critical and neutral trials. Follow-up logistic regressions for the critical and neutral trials separately indicated that the presence of human peers significantly reduced judgment accuracy on the critical trials, log-odds = -1.64, $SE = 0.30$, $z = -5.46$, $P < .00001$. No such effect was present for the robot peers, log-odds = 0.26, $SE = 0.37$, $z = 0.71$, $P = .48$. For the neutral trials, there were no significant differences between the conditions: control-human, log-odds = -0.30, $SE = 0.31$, $z = -0.97$, $P = .33$; control-robot, log-odds = -0.03, $SE = 0.32$, $z = -0.09$, $P = .93$. No other effects approached significance, $P > .91$. Accuracy patterns can be found in Fig. 2A. We also found that in the human-peer condition, 83% of the incorrect responses were the same as the confederate response ($\chi^2(1) = 15.114$, $P < .001$), indicating that participants were indeed conforming to the group response (Fig. 3).

This replicates the classical findings of Asch (*20–22*) and confirms recent studies (*33*). Im-

portantly, the drop in judgment accuracy with human peers was present exclusively for the critical trials, suggesting that the performance drop is not due to domain general anxiety driven by the presence of peers.

FIGURE 2 ABOUT HERE

Children

Adults do not appear to normatively conform to the humanoid robots used in the study, providing a challenge to the CASA hypothesis. However, since children are known to be more susceptible to social influence (*29, 30, 34, 35*), we evaluate this finding with young children in Experiment 2. Given the practical challenges of experiments using the original Asch paradigm involving child confederates, we focused exclusively on the influence of humanoid robot peers (cf. Section Outlook).

Participants ($N = 43$, 22 female, age: range = 7 – 9 years, $M = 8.5$ years, $SD = 0.5$) were randomly assigned to either the control ($n = 21$) or robot-peer ($n = 22$) condition. The methods and materials were identical to those from Experiment 1, with the exception that children were tested at school, rather than in a university lab.

We measured children's performance at the task when alone and when in the presence of robots using a 2 (condition: control vs. robot peer, between subjects) $\times$ 2 (trial type: critical vs. neutral, within subjects) experimental setup.

Analysis of Logistic Regression model

The analysis revealed that children are significantly influenced by the presence of robot peers (significant interaction between the two factors, condition and trial type, $\chi^2(1) = 11.1$, $P = .0009$). An analysis of the critical and neutral trials separately indicated that line judgment accuracy was lower in the robot-peer condition than in the control condition for critical trials

(log-odds = -0.37, $SE = 0.12$, $z = -3.17$, $P = .002$) but not the neutral trials (log-odds = 0.21, $SE = 0.15$, $z = 1.4$, $P = .16$). No other effects approached significance (all $P's > .30$). Accuracy patterns can be found in Fig. 2B and Table S1. We also found that in the robot-peer condition, 74% of the incorrect responses during the critical trials were identical to the responses provided by the robots ($\chi^2(1) = 14.785$, $P < .001$), again suggesting that conformity to the majority was taking place (Fig. 3).

FIGURE 3 ABOUT HERE

Discussion

It appears that adults in our study do not conform to the group of robots, confirming recent studies (*33*). Brandstetter et al. used four Nao humanoid robots to investigate informational and normative social influence in adults. The robots in their experiment were individualized with outfits and played pre-recorded human voices in order to focus on the appearance of the robots. Their setup also differed to ours in the length, presentation and number of stimuli. In 33 trials, Brandstetter et al. projected the lines of length up to $110cm$ onto a projection area and found that adult participants were influenced by their peers but not by the robots (neither with ambiguous nor unambiguous stimuli).

Children in our study on the other hand seem to conform to the robots. An alternative explanation for the findings is that children were not influenced or conforming, but rather that the relative novelty of the situation led to an overall decrease in judgment accuracy. This criticism holds no ground, as there was no accuracy decrease for the neutral trials. In fact, if anything, children performed slightly better for such trials (although this finding was not statistically significant), again indicating that they followed the suggestions made by the robots.

There is also the possibility that children were conforming to the robots' responses due to the authority invested in the robots by the adult experimenter. Even so, this still suggests that

the robots exert peer pressure and does not invalidate the observations and conclusions. Robots are likely to be owned by someone, people or organizations, and might as such be proxies for indirect social peer pressure.

The results of these experiments have both theoretical and practical implications. From a theoretical perspective, our results counter the notion that is central to the CASA hypothesis – that all people instinctively and automatically treat computer-based media as social (*17, 18*). While in certain tasks, adults do attribute human-like qualities to machines (*17*), they are capable of inhibiting the effects of normative influence, something which is not observed for human peers. We see this as a refinement of the CASA hypothesis, which impacts on the design of human-machine interaction in general.

Recent studies of online social networks have revealed that user behavior and decision making can be altered and manipulated through the selection of presented information (*36, 37*). Social robots are yet another social medium through which information may be transferred and communicated, and if trusted they can assert informational influence (*38*). The fact that robots have the power to induce conformity, even just in children, is relevant here and we believe our results are both timely and critical. In this light, care must be taken when designing the applications and artificial intelligence of these physically embodied machines, particularly as little is known about the long-term impact that exposure to social robots can have on the development of children and vulnerable sections of society (*39*). More specifically, problems could originate not only from intentional programming of malicious behavior (e.g. robots that have been designed to deceive) but also from the unintentional presence of biases in artificial systems (*40*) or the misinterpretation of autonomously gathered data by a learning system itself. For example, if robots recommend products, services or preferences, will compliance and thus convergence be higher than with more traditional advertising methods?

From a practical perspective, given that children do conform to erroneous suggestions made

by social robots, concerns are raised when using social robots with young people; while conforming can be beneficial (*41, 42*) (for example in health care or education), the potential for misuse or erroneous use cannot be ignored. This is a salient issue as there is a growing interest from the private/industrial sector in robots that interact with the general public and in particular with children. As this industrial market grows, so do the number of children potentially exposed to the issues outlined here.

A future in which autonomous social robots are used as aids for education professionals or child therapists is not distant. In these applications the robot is in a position in which the information provided can significantly impact the individuals they interact with. A discussion is required on whether protective measures, such as a regulatory framework, should be in place that minimize the risk to children during social child-robot interaction and what form they might take as not to adversely impact the promising development of the field.

Outlook

We conducted our experiment with children aged between seven and nine years. To create a more complete picture of conformity to robots, studies with different age groups, including older ages, need to be conducted such that the age ranges in which children and adults conform to robots can be determined.

Conducting the Asch experiment with children is difficult, as all but one of the children need to be confederates and convincingly act as fellow participants. Most studies on conformity with children have thus used a different paradigm to study conformity or used special optical setups giving the participant a different visual experience without the participant realizing (*35, 43*). A human-peer condition with children would have allowed a direct comparison between the results in the human peer condition and in the robot peer condition. The lack thereof, however, is a limitation of the current study.

A review of 133 Asch replication studies shows that conformity in adults has decreased since the 1950s (*28*). In addition, there is a correlation with a society's individualistic or collectivist nature. Compliance on the Asch paradigm is higher in societies with high collectivism, and it would be interesting to see if children and adults in collectivist cultures are more likely to yield to robots than individuals from individualistic cultures.

The sample sizes in our study are limited. Although sample sizes reflect commonly used sample sizes in the field, future studies could have more statistical power through using larger samples. With the current study, we can not study all possible factors impacting on conformity to robots. For instance we do not know how the robots are perceived by the participants or how participants judge the visual acuity of the robots. Allen argued that a greater similarity between the participant and the confederates will increase the likelihood of the participant perceiving the confederates as an appropriate reference group and hence will increase the level of conformity (*44*). Thus, adults might not form social bonds with small humanoid robots, but only with larger adult-size robots. Children on the other hand might not want to disagree with the robots for reasons that are as yet unexplored. All properties of design and behavior of the robots might potentially be factors that produce an influence on social conformity which need to be explored in future research.

Materials and Methods

We followed the experimental procedure as outlined by Asch (*20–22*) and used the same stimulus specification where possible (*22*). The adult experiments took place within a university lab setting while the experiments with the children were conducted at a local primary school in an empty classroom. Rather than presenting the stimuli on card, a TV screen was used. In the robot-peer condition software remotely orchestrated the response behavior of the three robots via a wireless network. The confederates, both human and robot, all followed the same pat-

tern of responses. All responses from participants and confederates were reported vocally and recorded by the experimenter using pen and paper. Participants (and confederates) were seated around a table, facing the TV screen (Fig. 1, B and C). For each of the 18 trials (12 critical, 6 neutral) the experimenter recorded the responses in a clockwise direction, beginning with the confederates and finishing with the participant. This order was constant for the human-peer and robot-peer conditions as was the seating plan. In the control condition no confederates were present.

Participants

60 adults took part in the experiment: 28 males ($M_{age} = 30.32$ years, $SD = 13.76$) and 34 females ($M_{age} = 31.48$, $SD = 14.61$). Participants were recruited via the online subject pool maintained by the School of Psychology at the University of Plymouth and were paid £4. They were randomly assigned to one of three conditions (control, robot peer, human peer), none of the participants were excluded (exclusion criterion: not using required vision correction). As participants were recruited through volunteer sampling, based on our one-way balanced between subjects design with three groups, the sample had a power level of .78 to detect a medium to large effect ($f = 0.4$) assuming an alpha level of .05.

43 children took part: 21 boys ($M_{age} = 8.47$ years, $SD = 0.58$) and 22 girls ($M_{age} = 8.50$, $SD = 0.50$). All were pupils at a local primary school in the Plymouth (UK) area and consent was obtained from both the school and parents. Children were pooled from one of two classes: Year 3 (aged 7 to 8, $n = 21$) and Year 4 (aged 8 to 9, $n = 22$). We have selected this age group as it is well-studied with respect to conformity, cf. (*45*), and younger children might not understand the task, as suggested by (*29*). Children were randomly assigned to either the control or robot-peer condition. Children would be excluded if they were not using required vision correction or if they felt uncomfortable. No children were excluded. The experimental sessions took place

over the course of a single school day and were located within a spare classroom within the school. No reward was provided, however at the end of the day a small presentation about robots was given by the experimenter. A power analysis showed that we had $> .71$ power to detect a medium to large effect ($d = .8$) assuming an alpha level of .05.

Materials

The length and order of the target and comparison lines were identical to the specifications outlined in original Asch studies (*20, 22*), see Table S2. A 32 inch LCD TV was used to display the stimuli as opposed to physical cards with printed lines. A laptop was connected to the screen running custom software to display the stimuli. In the human-peer condition the laptop's screen, only visible to the first confederate, also displayed the confederate answer allowing the first confederate to read this while looking at the TV screen. In the robot-peer condition this software was also used to orchestrate the behavior of the robots over a WiFi network.

The use of the TV screen introduced a deviation from the original Asch setup. We were unable to separate the target line and the matching comparison line by 40 inches (101.6 cm) as the TV screen was not wide enough for this. Instead we held this distance between the target line and the left hand comparison line constant at 40 cm. The horizontal distance between the edge of the screen and target line/right hand comparison line was 8.3 cm. All other dimensions were in accordance with the original experiments (*22*), see also Fig. S1 and Table S3. A smaller separation of target line and comparison lines makes the stimuli less ambiguous as it permits an easier comparison of line lengths, which should have no implications in studying normative social influence.

Three SoftBank Robotics Nao humanoid robots (Fig. 1A) were used as the confederates in the robot-peer condition. The Nao is a small 25 degree-of-freedom 58cm tall humanoid robot designed primarily for human-robot interaction. Each robot was autonomous, running custom

software that allowed it to be controlled by the software running on the experimenter's laptop. This software performed scripted behaviors that were run each time a new trial was displayed. The robots were seated at the table. In Experiment 1 they were seated on plastic boxes to elevate their position relative to the adult subjects (see Fig. 1C) to obtain approximately the same difference in face height between participant and robots across experiments. Only power cables were connected to the robots. The robots' head motor joint positions required to gaze at the TV screen, experimenter and participant were preprogrammed.

Procedure

Experiment 1

Subjects were randomly assigned to one of the three following conditions. In the 'control' condition the participants completed the task on their own, providing a baseline measure of performance. In the 'human-peer' condition the participants completed the task with three human confederates, serving as a replication of the original Asch experiments. In the 'robot-peer' condition the human confederates were replaced by robots.

Upon arrival in the experiment room, the confederates sat down in their agreed positions ensuring that the participant sat in the last seat (Fig. 1C). Participants (including the confederates) were briefed and consent was received. In the robot-peer condition, the briefing and obtaining of consent took place prior to entering the room. The robot's were already seated around the table when the participant entered.

Each participant was presented with an information sheet and a consent form. Participants were informed on the information sheet that they needed to perform a simple visual discrimination task in which they needed to indicate which of three comparison lines matched the length of a standard line in 18 such comparisons. They were also informed that all answers would be recorded on a prepared form.

An example visual stimulus was then used to provide a tangible instruction of the task. Participants were then offered the opportunity to ask for clarifications. Except in the control conditions, the experimenter defined the order of responses, clock-wise beginning with the first confederate. Following this the experiment began.

In the control condition participants performed the task alone, with only the experimenter in the room. In the human-peer condition the confederates provided their responses first. The first confederate was located opposite the participant, allowing the first confederate to see the laptop screen displaying the confederate answer while gazing toward the TV screen. All the other confederates followed her response. All robot confederates provided their response first as well.

Debriefing took place immediately after the experiment finished. Participants in the control condition were informed that they were in a control condition for the experiment. The nature of the experiment was also explained to them. Participants in the human- and robot-peer conditions were informed of the role of the confederates and what the aim of the experiment was: the measuring of normative social conformity. They also were given a questionnaire to collect demographic details, data on familiarity with and views of robots, and a personality test. All participants were requested to maintain confidentiality to avoid biasing future experiments.

Experiment 2

Experiment 2 mainly followed the same experimental procedure as described for Experiment 1. In Experiment 2, child subjects were only subject to the control and robot-peer conditions to which they were randomly assigned. Children were briefed while sitting at the table in the experiment room. Parental consent was obtained in advance. The children were not given any information sheet or questionnaire. The experimenter informed them orally that they needed to perform an “eye test” in which they needed to indicate which of three comparison lines

matched the length of a standard line in 18 such comparisons. They were also informed that all answers would be recorded on a prepared form. From here on, the course of the experiment was exactly the same as for the robot peer and control condition of Experiment 1, including the practice trial, the opportunity to ask for clarifications, the order of responses, and debriefing in the control condition. In the robot-peer condition, children were told during debriefing that the robots were trying to "trick" them and see whether they would agree with the robots. Children were also asked not to tell others about the experiment to avoid biasing future experiments.

Presentation of the robots

In the conditions where robots acted as confederates, the robots did not react to the participant when they entered and sat down. The experimenter outlined the instructions for the visual discrimination task and provided an example of the visual stimuli. When the lines were shown on screen the robots all gazed toward the experimenter as if listening to the instructions. The presentation of the real experimental trials commenced after this. From this stage onward, the scripted behavior of the robots was initiated each time the experimenter used the laptop to display the next set of comparison lines on the TV screen: all robots were instructed to gaze towards the screen, each with a different motor speed randomly selected uniformly from a given range. The robots paused for a random period between 0.75 and 1.5 seconds and then verbalised the desired response via an on-board text-to-speech engine. After giving a response, a robot occasionally looked at the participant for 1.5 seconds and then looked back at the screen. The purpose of this gaze behavior is to apply a certain amount of social pressure on the participants. A flow diagram of the scripted robot behaviour during the experimental trials can be found in Fig. S2.

A large part of this experiment depended on the manner in which the confederates were presented to the participant, particularly in the case of the robots. As such, care was taken

to present and treat the robots as individual social entities through the observable behavior, and how they were treated by the experimenter (i.e. the behavior of the experimenter directed toward the robots).

To provide the robots with a basic level of animacy, each robot was programmed to exhibit small behaviors to avoid the robot appearing static. Small motor movements were executed around the given gaze direction as were movements of the wrist joints and fingers. These motor commands were executed at random within a given time frame. Blinking behavior was also introduced through toggling power to the LED eyes at random intervals. Each of the robots was provided with an individual voice through altering the pitch of the text-to-speech engine. The eye colour of each robot was also individual. Fiducial markers were placed in the four outer corners of the screen, to allow to robot to see the screen.

Throughout the experiments, the experimenter's behavior toward the robots was as similar as possible to their behavior toward the participant. For example, during the task description, eye contact was made with both the participant and each individual robot. The robots were also given and referred to by names: Snap, Crackle and Pop.

In the robot-peer condition, adult subjects were informed in the information sheet that the aim of the research is to investigate visual discrimination in humans and robots and that each experiment involved 4 participants (a mixed group of humans and robots). Other than this, the reasoning for the robots being present was kept unspecified.

Ethics

The research design for this study was reviewed and approved by the Plymouth University Ethics Committee for the Faculty of Science and Engineering. Adult participants provided informed consent prior to the experiment and informed consent was provided by the parents of children prior to the experiments. Full debriefing in all conditions took place immediately after

the experiment ended.

Supplementary Material

Analysis of Logit (Logistic Regression) model.

Fig. S1. Specifications of visual stimuli presented to the participants.

Fig. S2. Flow diagram of the scripted robot behavior during the experimental trials.

Table S1. Discrimination accuracy across conditions.

Table S2. Specification of standard and comparison line lengths.

Table S3. Dimensions of the stimuli presentation.

Data S1. Text file of adult participant responses in Experiment 1.

Data S2. Text file of child participant responses in Experiment 2.

Further data

References

1. C. Breazeal, Sociable machines: Expressive social exchange between humans and robots, Ph.D. thesis, MIT, Artificial Intelligence Laboratory (2000).

2. T. Kanda, T. Hirano, D. Eaton, H. Ishiguro, *Human-computer interaction* **19**, 61 (2004).

3. T. Belpaeme, *et al.*, *Journal of Human-Robot Interaction* **1**, 33 (2012). In press.

4. W. Burgard, *et al.*, *Artificial intelligence* **114**, 3 (1999).

5. K. Wada, T. Shibata, *IEEE Transactions on Robotics* **23**, 972 (2007).

6. P. H. Kahn Jr, B. Friedman, D. R. Pérez-Granados, N. G. Freier, *Interaction Studies* **7**, 405 (2006).

7. F. Tanaka, A. Cicourel, J. R. Movellan, *Proceedings of the National Academy of Sciences* **104**, 17954 (2007).

8. C. Breazeal, *et al.*, *Topics in cognitive science* (2016).

9. J. Kennedy, P. Baxter, E. Senft, T. Belpaeme, *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, 2016), pp. 231–238.

10. T. Fong, I. Nourbakhsh, K. Dautenhahn, *Robotics and autonomous systems* **42**, 143 (2003).

11. P. H. Kahn, H. E. Gary, S. Shen, *Child Development Perspectives* **7**, 32 (2013).

12. G. F. Melson, *et al.*, *Journal of Applied Developmental Psychology* **30**, 92 (2009).

13. E. S. Kim, *et al.*, *Journal of autism and developmental disorders* **43**, 1038 (2013).

14. B. Scassellati, H. Admoni, M. Mataric, *Annual review of biomedical engineering* **14**, 275 (2012).

15. S. Thill, C. A. Pop, T. Belpaeme, T. Ziemke, B. Vanderborght, *Paladyn, Journal of Behavioral Robotics* **3**, 209 (2012).

16. M. Deutsch, H. B. Gerard, *The journal of abnormal and social psychology* **51**, 629 (1955).

17. B. Reeves, C. Nass, *How people treat computers, television, and new media like real people and places* (CSLI Publications and Cambridge university press Cambridge, UK, 1996).

18. C. Nass, Y. Moon, *Journal of social issues* **56**, 81 (2000).

19. C. Nass, J. Steuer, E. R. Tauber, *Proceedings of the SIGCHI conference on Human factors in computing systems* (ACM, 1994), pp. 72–78.

20. S. E. Asch, *Groups, leadership, and men* pp. 222–236 (1951).

21. S. E. Asch, *Readings about the social animal* **193**, 17 (1955).

22. S. E. Asch, *Psychological monographs: General and applied* **70**, 1 (1956).

23. C. Nass, Y. Moon, P. Carney, *Journal of Applied Social Psychology* **29**, 1093 (1999).

24. C. Nass, Y. Moon, N. Green, *Journal of applied social psychology* **27**, 864 (1997).

25. D. Abrams, M. Wetherell, S. Cochrane, M. A. Hogg, J. C. Turner, *British Journal of Social Psychology* **29**, 97 (1990).

26. T. F. Linde, C. Patterson, *The Journal of Abnormal and Social Psychology* **68**, 115 (1964).

27. A. H. Eagly, *Psychological Bulletin* **85**, 86 (1978).

28. R. Bond, P. B. Smith, *Psychological bulletin* **119**, 111 (1996).

29. M. B. Walker, M. G. Andrade, *The Journal of social psychology* **136**, 367 (1996).

30. M. Pasupathi, *Psychology and Aging* **14**, 170 (1999).

31. P. R. Costanzo, M. E. Shaw, *Child development* pp. 967–975 (1966).

32. T. J. Berndt, *Developmental psychology* **15**, 608 (1979).

33. J. Brandstetter, *et al.*, *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2014), pp. 1335–1340.

34. D. Haun, M. Tomasello, *Child development* **82**, 1759 (2011).

35. A. Hanayama, K. Mori, *Psychology* **2**, 661 (2011).

36. R. M. Bond, *et al.*, *Nature* **489**, 295 (2012).

37. A. D. Kramer, J. E. Guillory, J. T. Hancock, *Proceedings of the National Academy of Sciences* **111**, 8788 (2014).

38. N. Salomons, M. van der Linden, S. Strohkorb Sebo, B. Scassellati, *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, 2018), pp. 187–195.

39. N. Sharkey, A. Sharkey, *Interaction Studies* **11**, 161 (2010).

40. A. Caliskan, J. J. Bryson, A. Narayanan, *Science* **356**, 183 (2017).

41. J. Fasola, M. J. Mataric, *Proceedings of the IEEE* **100**, 2512 (2012).

42. J.-H. Han, M.-H. Jo, V. Jones, J.-H. Jo, *Journal of Information Processing Systems* **4**, 159 (2008).

43. K. Mori, M. Arai, *International Journal of Psychology* **45**, 390 (2010).

44. V. L. Allen, *Advances in experimental social psychology* (Elsevier, 1965), vol. 2, pp. 133–175.

45. D. B. Haun, van Leeuwen Edwin J.C., M. G. Edelson, *Developmental Cognitive Neuroscience* **3**, 61 (2013).

Acknowledgements: We thank Matthew Rule for acting as the human third confederate in the human-peer condition, Emily Ashurst for assisting with running the child experiments, and Paul Baxter for his insightful comments. **Funding:** This work was funded by the EU FP7 ALIZ-E (248116), FP7 DREAM (611391), FP7 Marie Curie Actions ITN RobotDoC (235065) and H2020 L2TOR (688014) projects, and grants from the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), Bielefeld University. **Author contributions:** DT

and ALV conceived the initial experiment with adults. DT, ALV, RR and TB designed and planned the experiments. ALV secured ethical approval. RR designed and developed the software for the laptops and robots. DT, ALV, and RR conducted the adult experiments where DT was the experimenter in all conditions. ALV and RR were confederates in the adult human-peer conditions. RR conducted the experiments with the school children. DT, RR and TB performed the data analysis. All authors contributed to the paper. **Competing interests:** The authors declare no competing financial interests. **Data and materials availability** All data needed to evaluate the conclusions in the paper are present in the paper or the Supplementary Materials.

Fig. 1. Overview of the experimental setup and visual stimulus. (**A**) The SoftBank Robotics Nao humanoid robot used as confederate. (**B**) Overview of the participant seating arrangement. In the control condition only the participant and experimenter were present. Participants' judgments are collected in a clockwise order beginning with the confederates and ending with the subject. (**C**) Illustration of the arrangement in a real setup. (**D**) Illustration of the visual stimuli presented to participants via a computer screen. The target line is located on the left and the three labeled comparison lines are located on the right. Participants say which of these matches the length of the target line.

Fig. 2. Discrimination accuracy across conditions. (**A**) The mean accuracy of the adults for the critical and neutral trials, across each experimental condition (control $n = 20$, robot peer $n = 20$, human peer $n = 20$). During the critical trials the presence of human peers leads to a significant decrease in discrimination accuracy due to subjects conforming with the human confederates. (**B**) The mean accuracy of the children during the discrimination task (control $n = 21$, robot peer $n = 22$, no human-peer condition). During the critical trials the presence of the robot-peers lead to a significant decrease in accuracy due to group conformity. Error bars denote 95% Confidence interval of the mean estimate; likelihood ratio test on logistic regression, * $P < .01$; ** $P < .001$.

Fig. 3. Breakdown of incorrect participant responses. The bars shows the ratio of conforming (i.e. going with the confederates' response) against non-conforming responses in the critical trials; for the adults in the human-peer condition ($n = 20$) and for the children in the robot-peer condition ($n = 22$). 83% of all incorrect responses from the adults were found to be conforming with the group of human confederates while children's conformity with the robots was 74%. Two-tailed χ^2 test, ** $P < .001$.

Spatial Referring Expressions in Child-Robot Interaction: Let's Be Ambiguous!

Christopher D. Wallbridge[1], Séverin Lemaignan[1], Emmanuel Senft[1], Charlotte Edmunds[1], Tony Belpaeme[1,2]

1 University of Plymouth
2 University of Ghent

*** christopher.wallbridge@plymouth.ac.uk**

Abstract

Establishing common ground when attempting to disambiguate spatial locations is difficult at the best of times, but is even more challenging between children and robots. Here, we present a study that examined how 94 children (aged 5-8) communicate spatial locations to other children, adults and robots in face-to-face interactions. While standard HRI implementations focus on non-ambiguous statements, we found this only comprised about 20% of children's task based utterances. Rather, they rely on brief, iterative, repair statements to communicate about spatial locations. Our observations offer strong experimental evidence to inform future dialogue systems for robots interacting with children.

1 Introduction

For children arriving in a new country, learning the language of their new home is an important part of their integration. Proficiency in the language of the host country is a vital condition for success at school. Even for children of migrants born in the host country, this may be an issue if the language used at school cannot be reinforced in the home. As tailored language classes are expensive and limited in time, we wish to explore if robot tutors can be used to complement language tutoring. This is encouraged by robots having been shown to be able to reduce anxiety in a second language learning when acting as a peer [1]. However there is still much to be considered when designing a robotic language tutor [5].

Figure 1. A child interacting with the robot in our study.

While most language tutoring systems focus on the learning of nouns and verbs, we wish to study the learning of spatial language instead: the vocabulary and grammatical constructions serving the communication of spatial relations. Spatial language is particularly challenging, as the semantics are often vague, context dependant and referent dependant. For example, in "the apple next to the bowl" the spatial referent "next" does not have boolean membership, but rather has a graded membership depending on the distance between objects and the size of the objects. A typical assumption in Natural Language Interaction Systems (NLIS) is that referring expressions (RE) are unambiguous descriptions of object locations and that a linguistic interaction between a user and a computer system follows a quite structured and clear interaction flow using unambiguous utterances [8]. This might be the case for spoken interfaces in banking systems or telephone ordering, but the

literature in socio-linguistics and dialogue systems show that language is much more dynamic than NLIS typically allows for, and this is specifically prominent in spatial RE.

Socio-linguistics suggests that people do not tend to use fully specified RE. Instead, they reduce the cognitive load by under-specifying the description and then rely on a strategy of repair to correct misunderstanding if necessary [7]. Rather than this being a one-way communication, it is a fundamentally social process. The person being addressed is expected to be an active contributor to the process of reaching *common ground*. Each participant in the conversation will contribute until a *grounding criterion* is met [6], i.e. when each contributor to the communication believes that they have understood enough for their current purpose. Pickering and Garrod [11] describe this partial alignment of common ground as the natural way in which we communicate. Full common ground is only necessary when there is difficulty reaching alignment.

Dialogue management systems have to take into consideration these under specified statements. One assumption that often made in interaction between two agents is that what is said by one, is how the other understands it. However this is not always true, even in human-human interaction [10]. Instead, continuous communication can allow a system to re-evaluate its belief state of the current environment, and the belief state of other communicative agents. For spatial tasks they are able to use contextual language to help with the positioning of an item [2]. Instead of complex statements that try to pinpoint the exact location in one sentence, a series of much simpler statements is used.

By contrast, implementations of RE generation and understanding for use in robotics often follow Gricean Maxims [9], such as the Incremental Algorithm [8]. These algorithms focus on a single statement that eliminates ambiguity. While communicating clearly and unambiguously about spatial references is one solution to the problem of communicating about space, more recent systems also incorporate perspective taking [12], which may alleviate the need for precise but verbose REs. With perspective taking we do see a more interactive approach. But this process still relies on reaching full alignment by eliminating ambiguity.

Our present study provides real-world data of children establishing common ground in the natural course of playing a game. We observed them either interacting with other children, with adults or with a robot using a Wizard of Oz setup. The study provides opportunities for the children to use a large set of spatial language, perspective taking and establishing a common point of reference, whilst being easy to replicate.

2 Study Design

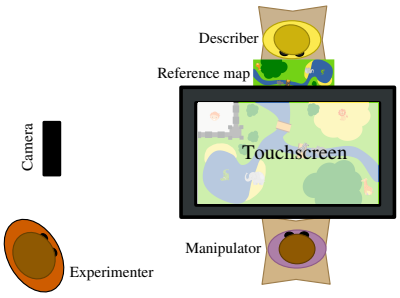


Figure 2.The experimental setup. A top down view showing the position of the manipulator and describer sitting opposite each other with the "Sandtray" screen in the middle. The experimenter is sitting to the side with a camera recording the participants.

We collected data from 94 children between the ages of 5 and 8. They were assigned to one of three conditions: child-child, child-adult or child-robot. For the child-child and child-adult conditions children from two different schools were used. They participated during the day at their school in a room for individual teaching. In the child-child condition two children from the same class participated together. In the child-adult condition a child participated with an experimenter. Those in the child-robot condition were recruited from register held by the Babylab at the University of Plymouth.

Following a sandbox paradigm [3], one child and a partner (child, robot or adult) are sitting on opposite sides of a large touchscreen (Fig. 2). The screen presents a background with different areas: a castle, a desert, two rivers with bridges, a lake, two beaches and many bushes or trees.

One agent, hereafter called the *describer*, has to guide the other agent, called the *manipulator*, to move items on the touchscreen to a desired location. The describer is provided with a reference map, which is kept hidden from the manipulator, with the desired position of eight items (Fig. 3).

While it has been shown that pointing can influence the words used [13], the task could be easily completed without words if gestures were allowed. As we were focused on the language being used, the describer was instructed not to use pointing gestures. If children attempted to use pointing they were reminded that this was not allowed.

Figure 3.An example of the reference map given to a child to describe. The eight items (face, crocodile, elephant, zebra, hippo, lion, giraffe and ball) are shown in the desired location that they need to be moved to. The child describes the position on his map for an agent to manipulate into the correct position.

The touchscreen presents a background with different areas (Fig. 3). Eight movable items have to be moved to specified locations on the map. The reference maps were designed to elicit a number of different ways to describe the position of objects. Some objects were facing a particular direction, to encourage locutions like 'in front of' or 'behind'. Features, such as the bridges and bushes, were repeated so as to require disambiguation. Verbal disambiguation was also elicited by the relatively small size of the screen, which limits the effectiveness of joint gaze to identify the correct location for an object.

In the case of the child-child and child-adult conditions, after the first map was completed, the role of manipulator and describer would be swapped. In the case of the child-robot condition the child would be invited to describe the second map. The robot itself would appear to move objects around the touchscreen via the use of a Wizard of Oz control interface, held by an experimenter. The experimenter is able to move an object on their interface, the robot would then move its hand to point at the object and then move its hand to point at the target location, with the object moving with it.

3 Results

For statistical power reasons, we focused our current observation of results on the child-child interaction (Child-Child=60, Child-Adult=26, Child-Robot=8), while providing more qualitative observations of the other conditions in the discussion.

We observed an average of 7.12 (SD=7.50) repair statements used per round (one round consisted of one map with eight objects to be moved). The SD shows large inter-personal variations. There were comparatively few cases of repair statements requiring spatial perspective taking (M=0.56 per round). Despite being told not to use them, there was an average of 2.43 (SD=3.03) pointing gestures used per round.

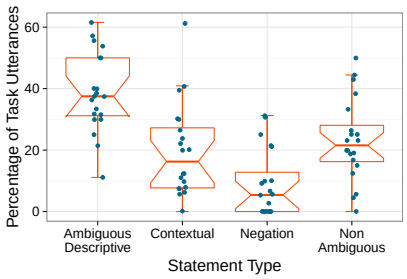


Figure 4. Break down of on-task statements. Ambiguous descriptive statements were a significantly higher proportion than the other statement types.

We took all the on-task statements from a sample of 10 child-child sessions, giving us data from 20 children. The statements were divided into the following categories: Ambiguous-Descriptive (statement refers to more than one location e.g.'the zebra is on a bridge'), Contextual (statement following from previous statements, that would make no sense to a third person entering the conversation e.g. 'the other one'), Negation(statement indicating that it is an incorrect location with no further description e.g. 'no'), Non-Ambiguous (statement that describes only one possible location e.g. 'the crocodile is in the big lake') and Pointing.

On average Ambiguous-Descriptive statements were used 38.6% of the time, Contextual in 13.1%, Negation in 9% and Non-Ambiguous in 23.2%. Using a Welch

two-sample t-test we find that the Ambiguous-Descriptive statements are used significantly more than any other type of statements, and Cohen's d test shows a large effect size in each case (Contextual: $t(38) = 4.2$, $p < .001$, $d = 1.34$; Negation: $t(38) = 7.8$, $p < .001$, $d = 2.48$; Non-Ambiguous: $t(38) = 3.7$, $p < .001$, $d = 1.17$).

4 Discussion

Our observations show that interactions between children (and between children and robots) are highly dynamic, fast-paced and relying on the situatedness and embodiment of the conversation partners [4], very unlike the "walkie-talkie exchanges" typically used in Human-Robot Interaction. Between children, as soon as the manipulator has enough information to make a guess they will often start moving the objects, without waiting until enough information is given as to be non-ambiguous. This has two possible outcomes: either they guess right, or it causes the describer to generate a repair statement. It also appears that typically it is easier for the describer to let the manipulator start moving the objects – knowing that the position they described is ambiguous – so that they may then generate a short, easily understood, repair, reducing the cognitive load. In fact we see that the robot's inability to change course after it has started moving an object caused frustration to the child describing.

In the child-robot condition there appeared to be a reduction of the repair statements when the robot moved items incorrectly. This could be caused by many factors, such as the children feeling more nervous with the robot, the expectations they have of its abilities and the absence of some basic social cues, such as back channelling and lack of eye contact, all of which made the interaction laborious.

Pointing was still prevalent, despite it being disallowed and discouraged (even the experimenter was found pointing or indicating directions). Future work could look at a different methodology to encourage the combination of gestures and language.

5 Conclusion

Counter to many implementations that seek to eliminate ambiguity entirely, we find that children tend to use many ambiguous statements when describing the location of objects. As such *the robot, when being given RE, must expect ambiguous statements.* It should not wait for further information, but rather start acting on the information it has, as this will also assist in the process of description. This also means that the robot should be prepared to react quickly to repair statements by enabling it to diverge from its current action to take into account the new information.

This also means *the robot should be allowed to be ambiguous in its descriptions.* This may be beneficial to reduce processing requirements for the robot itself, but also may help reduce the cognitive load for its conversational partner. When doing so, the robot should monitor closely the reaction of its partner, and be prepared to provide timely repairs to lead the implicit, interactive disambiguation process.

Our next steps are to implement a more interactive robot to collect more data with children interacting with the robot. Using this data we will be able to build an effective framework for natural spatial communication between children and robots.

6 Acknowledgements

This work was supported by the EU H2020 L2TOR project (grant 688014), the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227) and the EU FP7 DREAM project (grant 611391).

References

1. M. Alemi, A. Meghdari, and M. Ghazisaedy. The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics*, pages 1–13, 2015.

2. T. Baumann, M. Paetzel, P. Schlesinger, and W. Menzel. Using Affordances to Shape the Interaction in a Hybrid Spoken Dialog System. In *Proceedings of ESSV*, Bielefeld, Germany, Mar. 2013.

3. P. Baxter, R. Wood, and T. Belpaeme. A touchscreen-based'sandtray'to facilitate, mediate and contextualise human-robot social interaction. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 105–106. ACM, 2012.

4. T. Belpaeme, S. J. Cowley, and K. F. MacDorman. *Symbol grounding*, volume 21. John Benjamins Publishing, 2009.

5. T. Belpaeme, P. Vogt, R. van den Berghe, K. Bergmann, T. Göksun, M. de Haas, J. Kanero, J. Kennedy, A. C. Küntay, O. Oudgenoeg-Paz, et al. Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*, 2017.

6. H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive science*, 13(2):259–294, 1989.

7. H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

8. R. Dale and E. Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263, 1995.

9. H. P. Grice, P. Cole, J. Morgan, et al. Logic and conversation. *1975*, pages 41–58, 1975.

10. G.-J. M. Kruijff, M. Janíček, and P. Lison. Continual processing of situated dialogue in human-robot collaborative activities. In *RO-MAN, 2010 IEEE*, pages 594–599. IEEE, 2010.

11. M. J. Pickering and S. Garrod. Alignment as the basis for successful communication. *Research on Language & Computation*, 4(2):203–228, 2006.

12. R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken. Which one? grounding the referent based on efficient human-robot interaction. In *RO-MAN, 2010 IEEE*, pages 570–575. IEEE, 2010.

13. A. Sauppé and B. Mutlu. Robot deictics: How gesture and context shape referential communication. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 342–349. ACM, 2014.

KOÇ UNIVERSITY

How do robot gestures help second language learning?

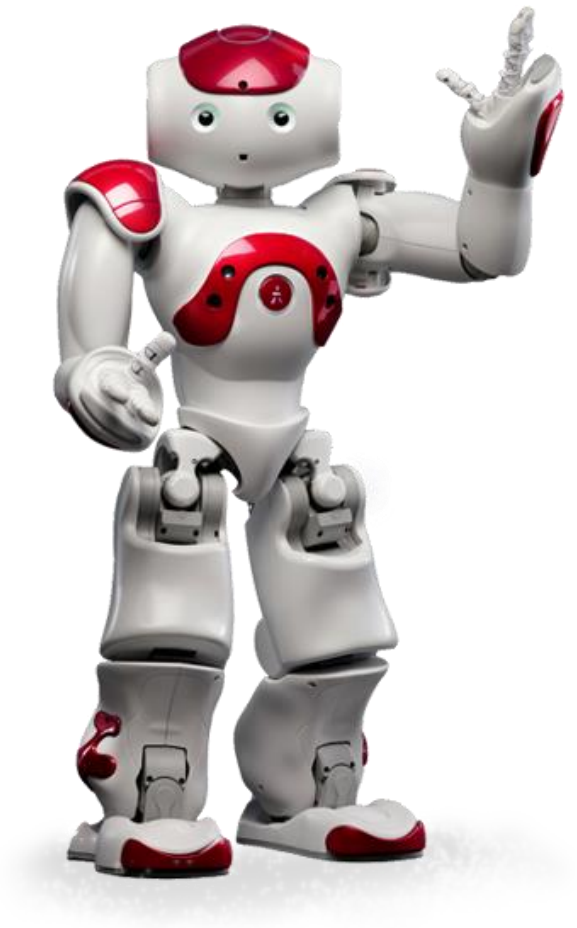
Junko Kanero, Özlem Ece Demir-Lira, Sümeyye Koşkulu, Cansu Oranç, Idil Franko, Aylin C. Küntay, & Tilbe Göksun

L2TOR

Grant agreement no. 688014

European Commission | Horizon 2020 European Union funding for Research & Innovation

Unique Strengths of Robot Learning Companions



- Can act as friendly learning partners
- Can use different languages
- Can perform gestures
  - Gestures facilitate language learning in children
    (e.g., Hostetter, 2011; Sueyoshi & Hardison, 2005; Valenzeno et al., 2003)

Under what conditions are gestures beneficial for young learners?

Common Practice in Educational Robotics

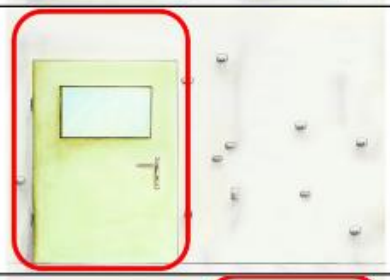
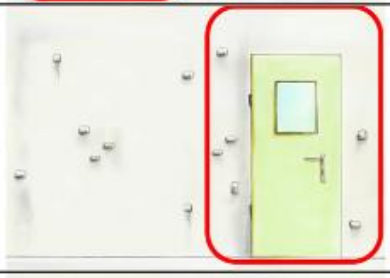
- Gestures are created not by gesture experts but by roboticists
- In some cases, an additional study to make sure gestures are “good enough”
- But is that really enough?

Today's Talk

- RQ1: *Can children learn L2 words from a robot tutor that is performing gestures?*
- RQ2: *Are some robot gestures better than other gestures?*
  - Study 1: *Types of gestures*
  - Study 2: *Match between words and gestures*

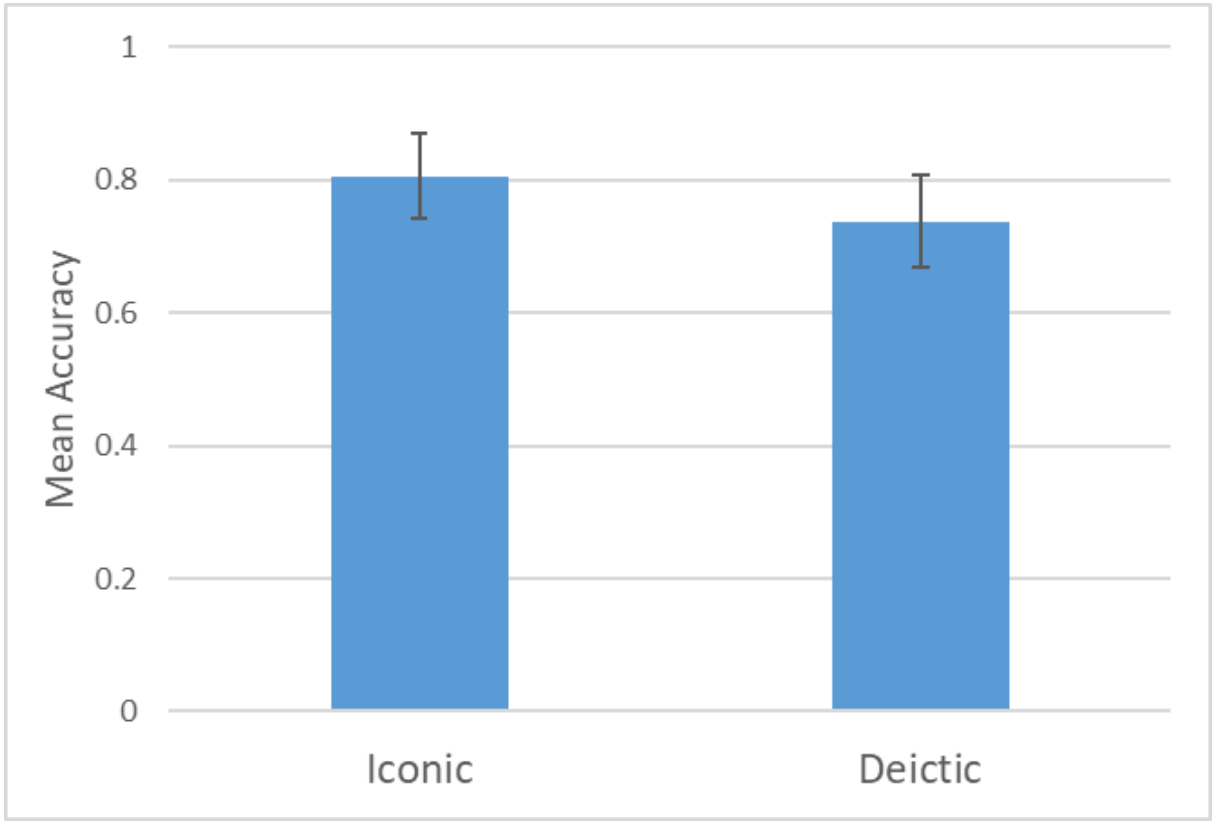
Study 1: Types of Gestures

- 88 Turkish-speaking preschoolers
- One-on-one lesson using a tablet
- 6 pairs of English adjectives
  - big-small, tall-short, high-low, wide-narrow
- Human tutor vs. Robot tutor
  - Iconic Gesture – represents the meaning of the word
  - Deictic Gesture – indicates the reference of a word
  - On-Screen Highlighter

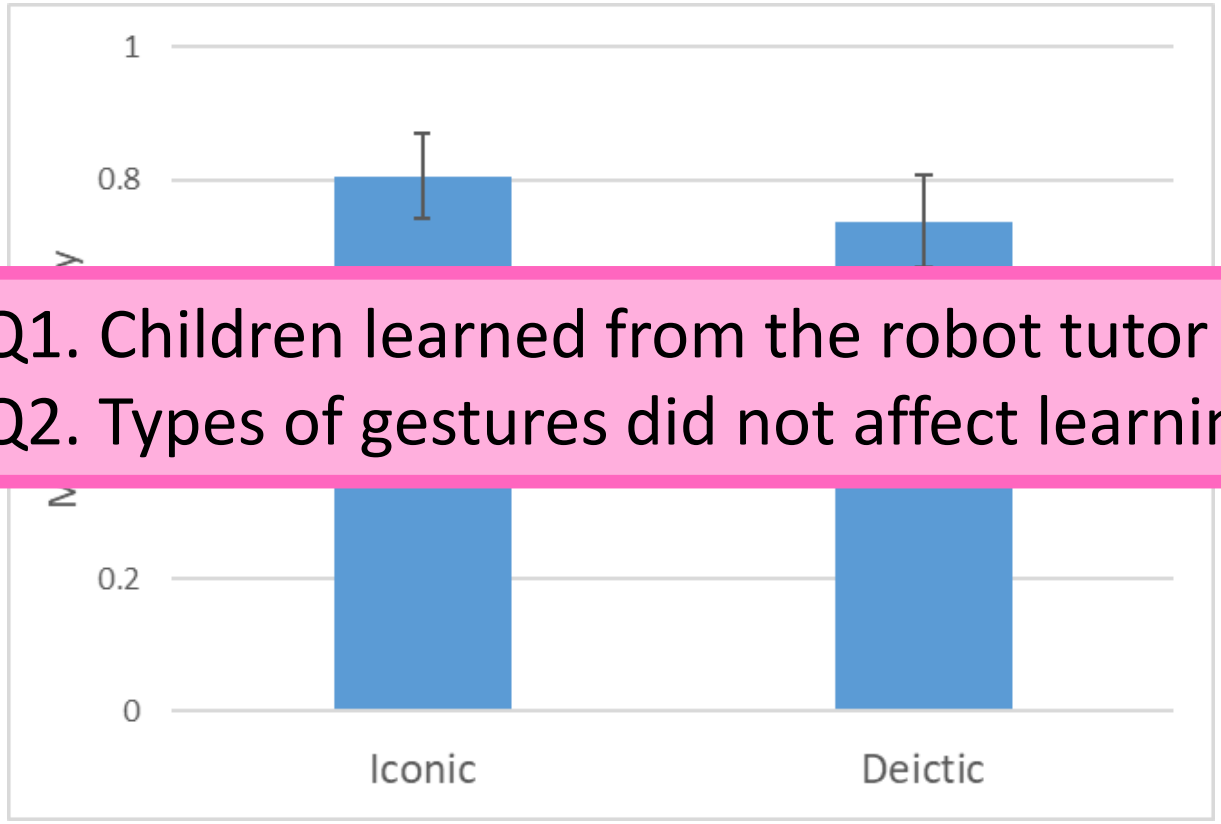
	Gesture (Iconic or Deictic)	On-Screen Highlighter
Noun Learning		
Adjective Learning		
		
Test		

Example: Iconic Gesture Trials

Iconic Gesture vs. Deictic Gesture
(Robot Condition)

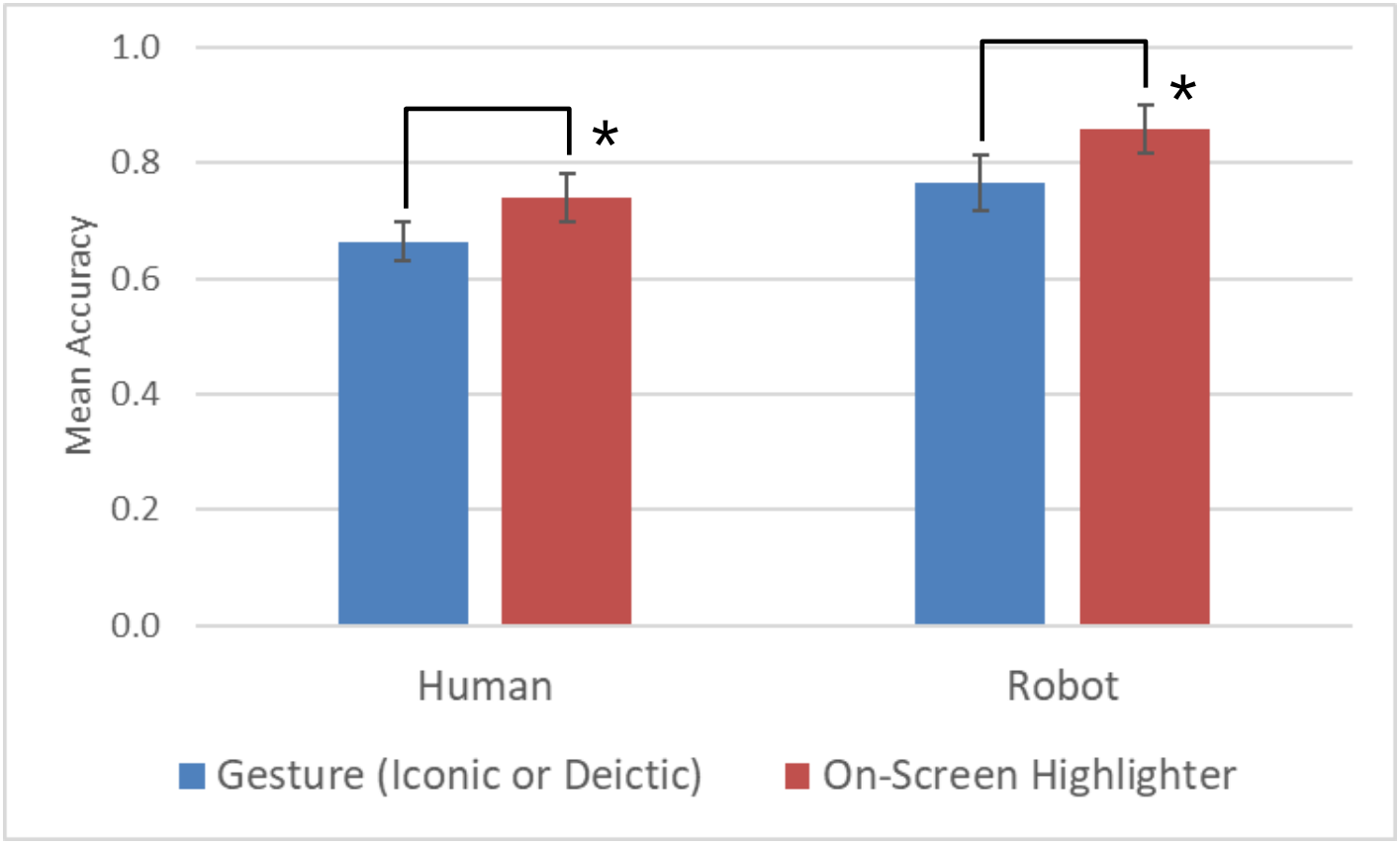


Iconic Gesture vs. Deictic Gesture
(Robot Condition)

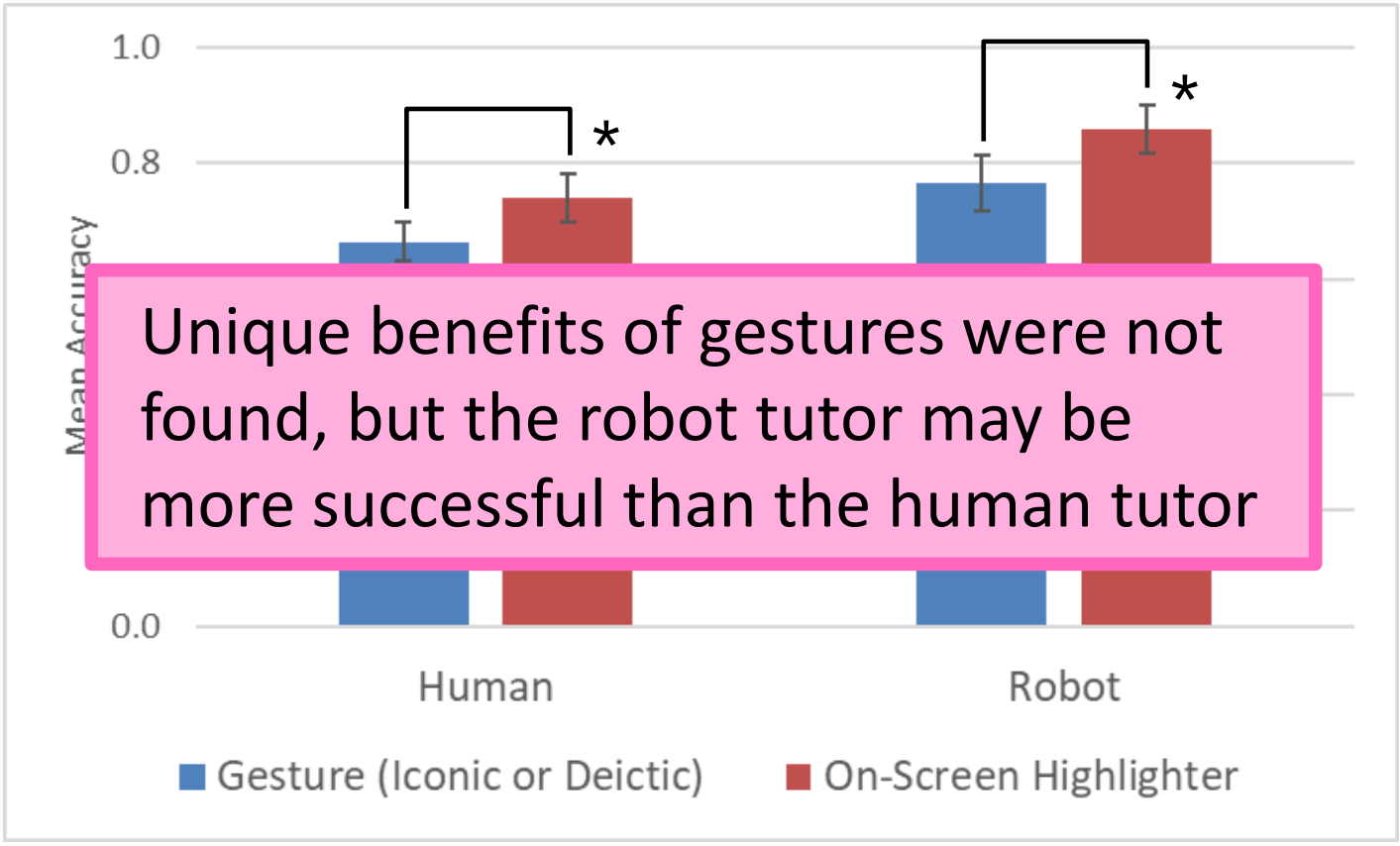


RQ1. Children learned from the robot tutor
RQ2. Types of gestures did not affect learning

Gesture vs. On-Screen Highlighter



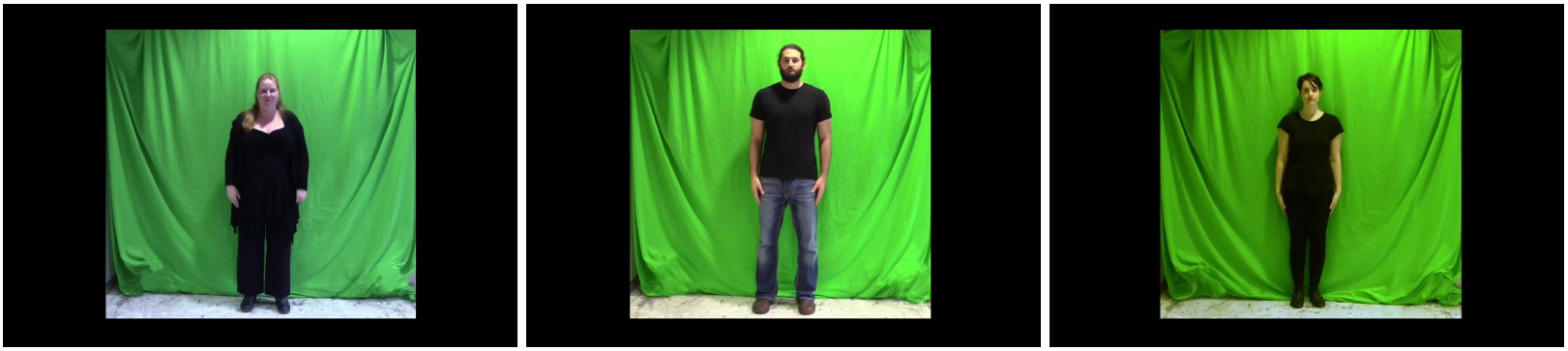
Gesture vs. On-Screen Highlighter



Unique benefits of gestures were not found, but the robot tutor may be more successful than the human tutor

Study 2: Word-Gesture Match

- **<u>Step 1: Production Task with Adults</u>**
  - 3 English-speaking adults performed gestures for 53 English words



- **<u>Step 2: Rating Task with Adults</u>**
  - 30 English-speaking adults rated how well the gestures represented the words

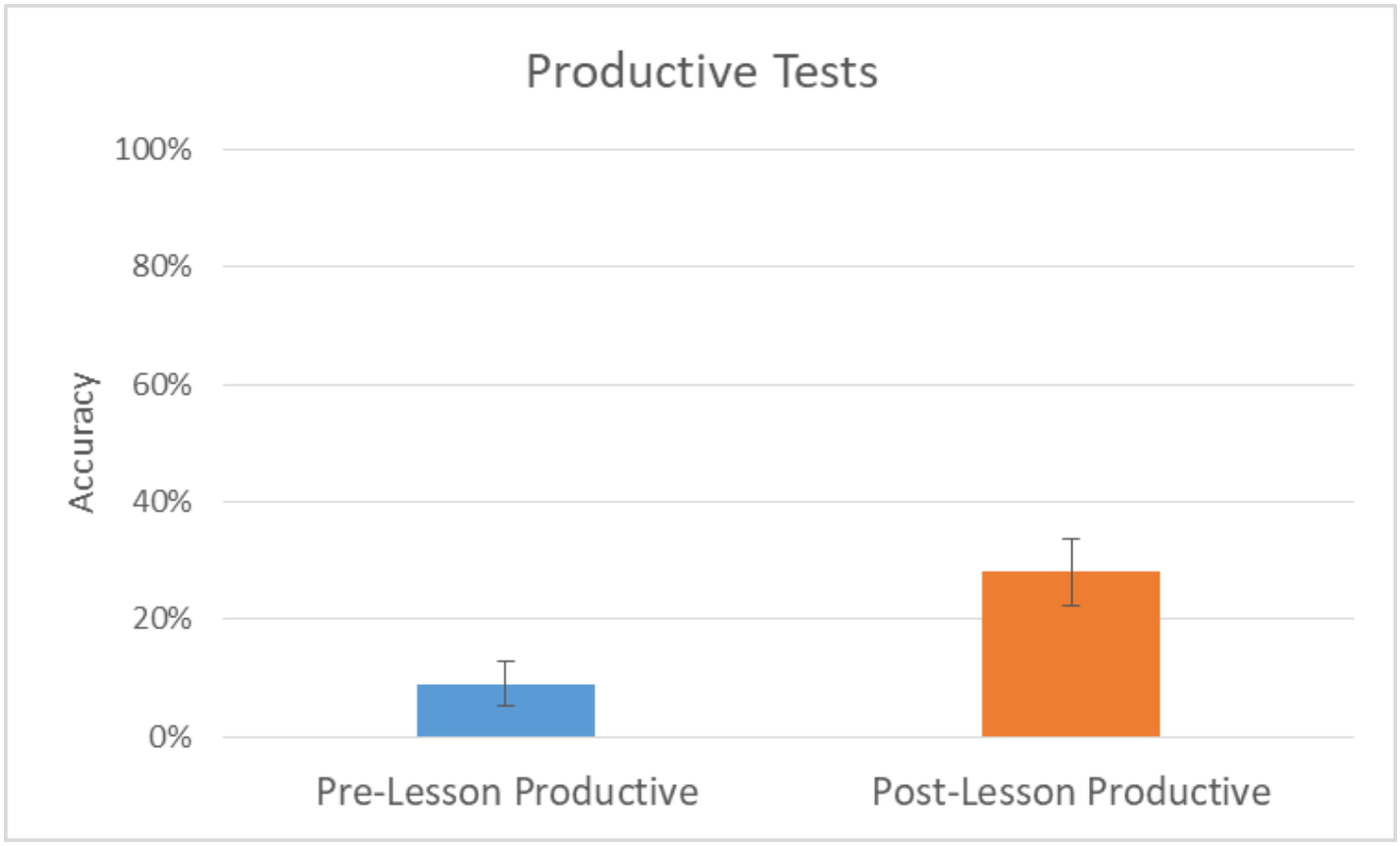
Study 2: Word-Gesture Match

- **Step 3: Learning Task with Children**
  - 20 Turkish-speaking preschoolers (*M* = 65.33)
  - One-on-one lesson without a tablet
  - 5 English verbs
    - *sliding*, *falling*, *climbing*, *walking*, and *throwing*

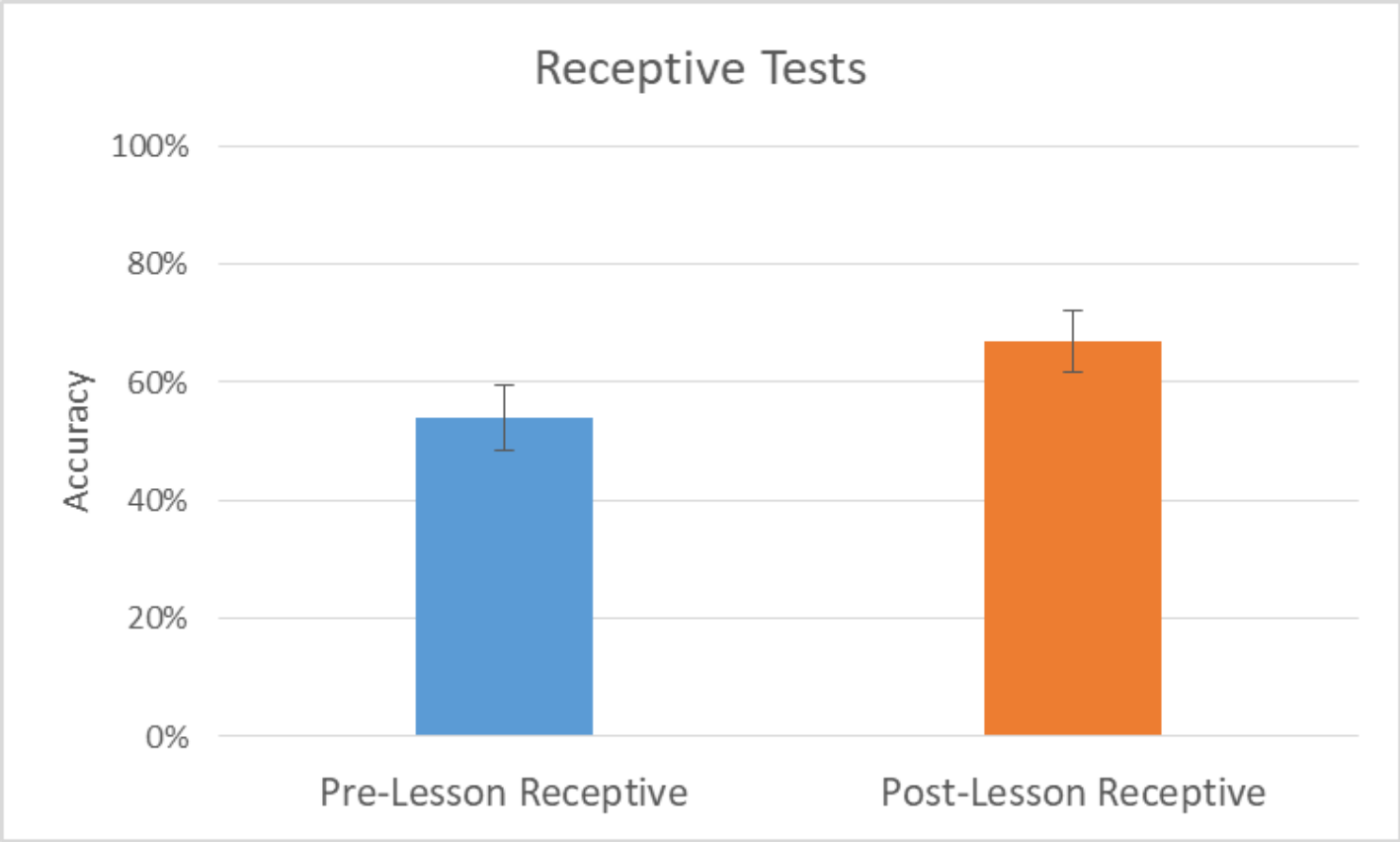
Final Set of Gestures

Experimental Setting

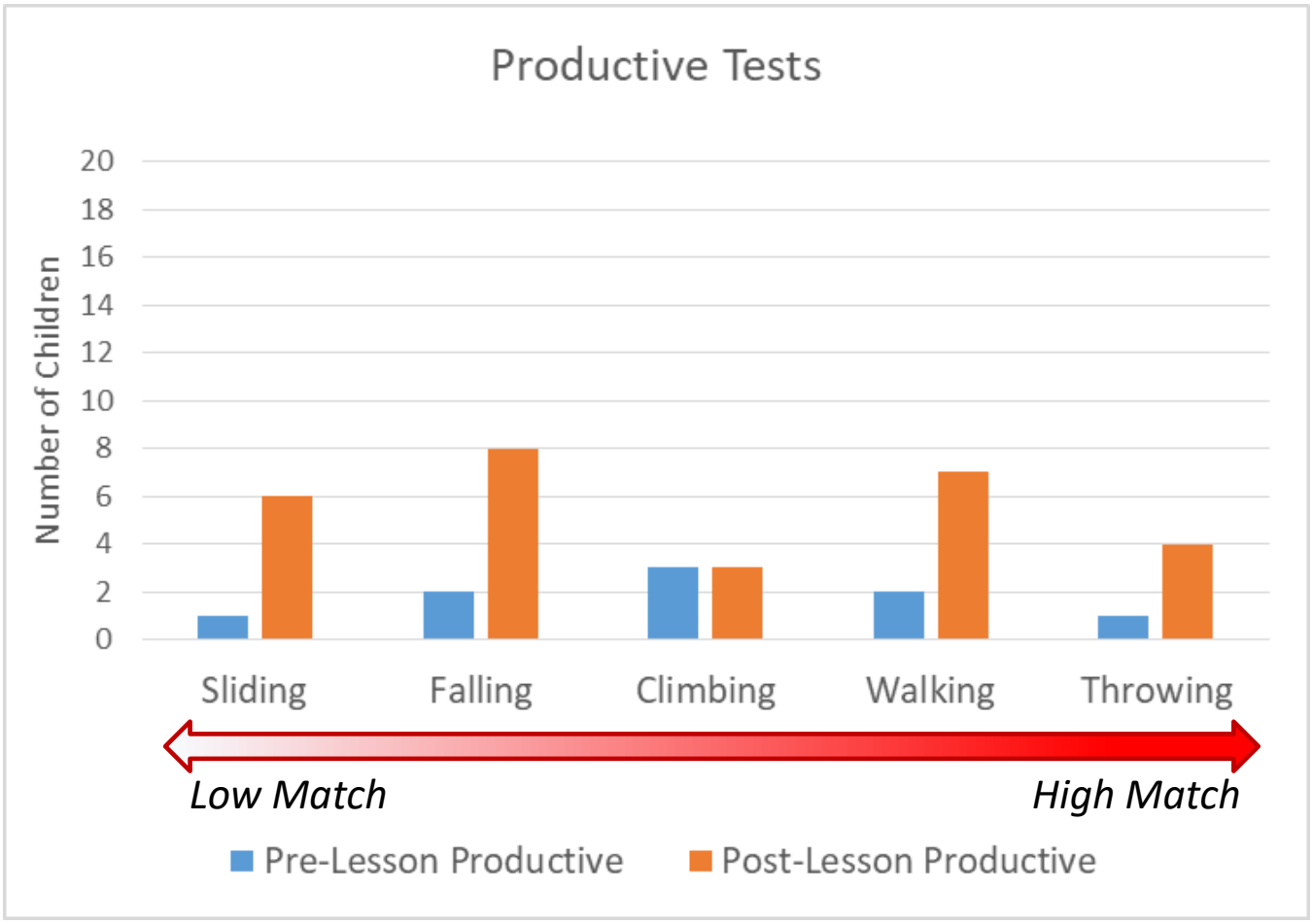
RQ1: Did children learn?



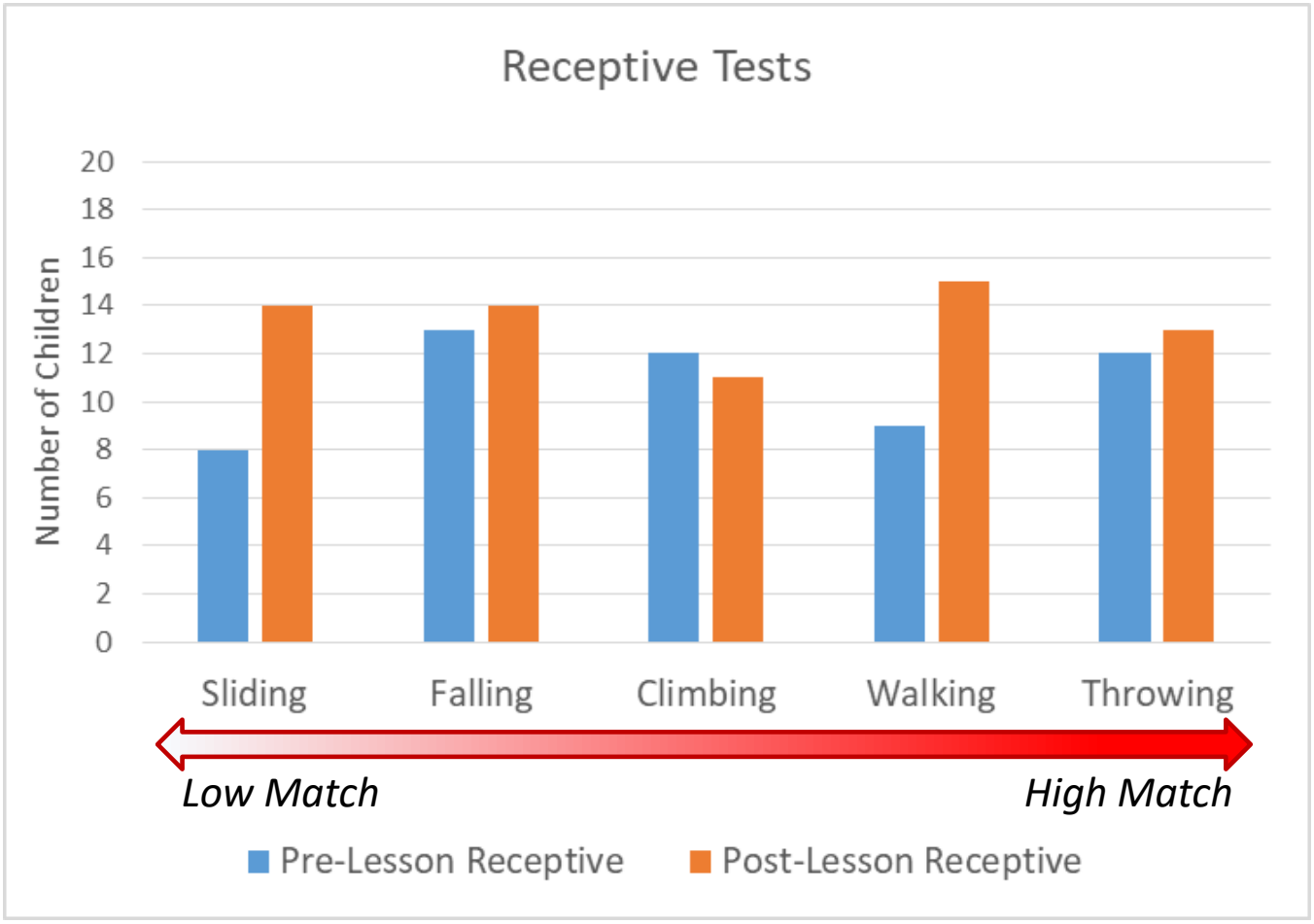
RQ1: Did children learn?



RQ2: Match between word and gesture?



RQ2: Match between word and gesture?



Discussion

- Children can learn L2 words from the NAO robot
- Types of gestures nor match between the word and gesture did not seem to matter
- Study 1 vs. Study 2
  - Tablet vs. No tablet
  - Elementary vs. Advanced Vocab
  - No translation vs. Translation

Discussion

- "Good enough" gestures?
  - The contribution of gestures may be simply drawing attention and engaging the child
  - Can even be destructive in some situations? (e.g., when children need to focus on screen)

KOÇ UNIVERSITY

Teşekkür ederim!
(Thank you!)

L2TOR

European Commission | Horizon 2020 European Union funding for Research & Innovation

Robotların Jestler ve Geri Bildirim ile Okul Öncesi Dönemde İngilizce Eğitimine Katkıları

Sümeyye Koşkulu, Junko Kanero, Cansu Oranç, Tilbe Göksun, Aylin C. Küntay

Sosyal robotlar, okul öncesi dönemde çocukların ikinci dil eğitimine destek olması amacıyla sıklıkla kullanılmaya başlanmıştır (Kanero va., 2018). Ancak robotların hangi özelliklerinin bu eğitime anlamlı katkıda bulunabileceğine dair yeterli çalışma yoktur. Bu araştırma, robotların sözlü (geri bildirim) ve sözlü olmayan (jest) davranışlarının İngilizce sözcük öğrenimine etkilerini incelemektedir.

Çalışma 1'in amacı, robot jestlerinin ve bu jestlerin sözcükleri temsil etme derecesinin İngilizce öğrenimine katkılarını araştırmaktır. Çocuklara 5 İngilizce fiil insansı robot NAO tarafından kelimeleri farklı derecelerde temsil eden ikonik jestler uygulanarak öğretilmiştir. Çocukların kelime bilgileri dersten önce ve sonra alıcı ve ifade edici testler aracılığıyla ölçülmüştür. Veri toplama süreci devam eden çalışmaya 4.5-6.5 yaş aralığında 20 çocuk katılmıştır. Sonuçlar çocukların alıcı dil skorlarında anlamlı bir artış olduğunu gösterirken ($t(19) = -2.89$, $p = .01$), ifade edici dil skorlarında sınırda anlamlı bir artış olduğunu göstermiştir ($t(19) = -1.72$, $p = .10$). Alıcı ve ifade edici dil skorları için yapılan analizler ise jest-fiil temsil derecesinin çocukların kelimeleri öğrenmesi açısından farklılık oluşturmadığını göstermiştir.

Çalışma 2'nin amacı insanların verdiği sözel geri bildirimin etkilerinden yola çıkarak robot tasarımını bilgilendirmektir. Bu amaçla, Çalışma 1'den üç İngilizce fiil çocuklara tanıtılmıştır. Daha sonra söylenen fiili ekrandaki üç animasyon arasından göstermesi istenen çocuklar üç gruba ayrılmış ve yanlış seçeneği göstermeleri durumunda her gruba farklı sözel geri bildirim verilmiştir: (1) İngilizce cümlenin tekrarı, (2) Seçilen yanlış seçeneğin isimlendirilmesi, (3) Doğru seçeneğin gösterilmesi. Son olarak, çocuklara farklı animasyonlar gösterilerek öğrendiklerini aktarmaları istenmiştir. Çocukların alıcı kelime bilgileri ölçülmüştür. Çalışmaya 3-6 yaş aralığında 78 çocuk katılmıştır. İngilizce bilgisi düşük olan çocukların (2) numaralı koşulda (1) numaralı koşula göre kelimeleri daha iyi öğrendikleri görülmüştür, $F(4, 40) = 2.488$, $p = .059$.

Bu bulgular robotlar tarafından yürütülen İngilizce derslerinde jestlere ihtiyaç duyulduğunu ve jestlerin dolaylı yoldan çocuğun dikkatini çekme ve sürdürme gibi roller oynadığını ortaya koymaktadır. Ayrıca çocukların İngilizce kelime bilgisi onlara verilen sözel geri bildirimlerden yararlanma biçimlerini etkilemektedir.

Anahtar sözcükler: sosyal robotlar, yabancı dil eğitimi, jestler, geri bildirim

Using gestures in L2 vocabulary teaching: Human or robot tutors?

Ece Demir-Lira[1,2], Cansu Oranç[1], Junko Kanero[1], Sümeyye Koskulu[1], İdil Franko[1], Zeynep Adıgüzel[1], Tilbe Göksun[1] & Aylin C. Küntay [1]

[1]Koç University, [2]University of Iowa

Introduction

- Gestures facilitate language learning (e.g., Hostetter, 2011; Novack & Goldin-Meadow, 2015). However, not all studies observed the facilitatory effects of gestures, and specific conditions under which gestures aid language learning remain under debate (Congdon et al., 2018).
- Robot's ability to gesture suggested as a strength over other technological tools (e.g., Kanero et al., 2018).

Study goals

- We examined whether and how gestures facilitate second-language (L2) vocabulary learning in children by varying the:
  - Type of facilitatory tool (gesture, on-screen highlighter)
  - Gesture type (deictic, iconic)
  - Tutor (human, robot)

Method

Participants

- 5- and 6-year-old Turkish-speaking children
  - Human tutor: n = 41, *M* = 66.9 months, 22 Females
  - Robot tutor: n = 37, *M* = 69.9 months, 21 Females

Measures

- 8 English measurement words: Big-small, long-short, wide-narrow, high-low
- Prior knowledge of words in Turkish, but not English

Procedure

- Children interacted with Human or Robot tutor.
- Each child experienced highlighter + 1 of the 2 gesture conditions (4 words per condition, 3 blocks)
  - On-screen highlighter: no gesture, a red rectangle flashed around the object to draw attention. (Figure 1a)
  - Iconic Gesture: tutor produced an iconic gesture (Figure 1b)
  - Deictic Gesture: tutor pointed (palm-hand) to object on screen (Figure 1c)
  - Images of objects representing words present on screen in all conditions (see Figure 1)
- During test, children asked to point to the object corresponding to the target measurement word.
- Generalization of measurement words to new objects assessed at the end of the session.

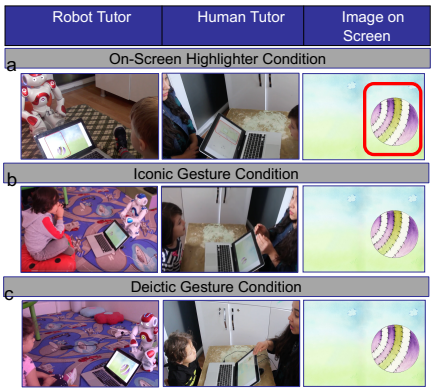


Figure 1. Study design

Results

Results for test questions

- Mixed-effects ANOVA using tutor type (human, robot), gesture type (iconic, deictic), sex (female, male) as between-subjects and facilitatory tool (gesture, highlighter) as within-subject variable on proportion correct
  - Significant effect of facilitatory tool: Highlight > Gesture, $F(1,74) = 6.764$, $p = .001$, partial eta-squared $\eta p^2 = .084$
  - Marginally significant effect of tutor: Robot > Human, $F(1,74) = 3.023$, $p = .086$., partial eta-squared $\eta p^2 = .039$

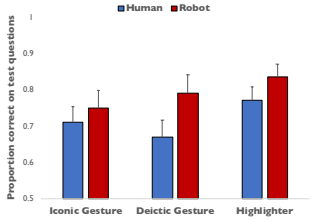


Figure 2. Proportion correct by tutor, gesture and facilitatory tool

Results on generalization questions

- No significant effect of tutor, gesture type, or facilitatory tool on performance, all *p*'s >.10

Discussion

- Gestures did not result in significantly better learning than an on-screen highlighter, and gesture type had no significant effect on learning outcomes.
  - Gestures may not be as effective for learners with no prior knowledge and when task requires attention to visuals in the learning environment (Congdon et al., 2018).
  - Significant difference between the gesture and on-screen highlighter conditions suggests that the role of gesture in word learning different than simply guiding attention (Novack et al., 2016).
- Children might learn L2 words better with a robot than with a human.
  - Possible explanations for the robot tutor advantage, such as novelty and the less lesson-like, friendlier atmosphere the robot might have created (Conti et al., 2017; Kanero et al., 2018).
- Future directions:
  - Remove visuals, provide translations, beat gestures

References

Conti, D., Di Nuovo, A., Cirasa, C., & Di Nuovo, S. (2017, March). A comparison of kindergarten storytelling by human and humanoid robot with different social behavior. In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (pp. 97-98).

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. Psychological bulletin, 137(2), 297.

Congdon, E. L., Kwon, M. K., & Levine, S. C. (2018). Learning to measure through action and gesture: Children's prior knowledge matters. Cognition, 180, 182-190.

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. Language and cognitive processes, 24(2), 313-334.

Kanero, J., Geçkin V., Oranç, C., Mamus, E., Küntay, A. C., & Göksun, T. (2018). Social robots for early language learning. Child Development Perspectives. doi: 10.1111/cdep.12277

Novack, M., & Goldin-Meadow, S. (2015). Learning from gesture: how our hands change our minds. Educational psychology review, 27(3), 405-412.

Novack, M. A., Wakefield, E. M., Congdon, E. L., Franconeri, S., & Goldin-Meadow, S. (2016). There is More to Gesture Than Meets the Eye: Visual Attention to Gesture's Referents Cannot Account for Its Facilitative Effects During Math Instruction. In Proceedings of the 38th Annual Meeting of the Cognitive Science Society.

Acknowledgements

1. *This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 688014.*
2. *We thank all schools, families and children who participated.*

Who Can Benefit from Robots? Effects of Individual Differences in Robot-Assisted Language Learning

Junko Kanero
Department of Psychology
Koç University
Istanbul, Turkey
jkanero@ku.edu.tr

Idil Franko
Department of Psychology
Koç University
Istanbul, Turkey
ifranko15@ku.edu.tr

Cansu Oranç
Department of Psychology
Koç University
Istanbul, Turkey
coranc14@ku.edu.tr

Orhun Uluşahin
Department of Psychology
Koç University
Istanbul, Turkey
oulusahin14@ku.edu.tr

Sümeyye Koşkulu
Department of Psychology
Koç University
Istanbul, Turkey
skoskulu17@ku.edu.tr

Zeynep Adıgüzel
Department of Psychology
Koç University
Istanbul, Turkey
zadiguzel13@ku.edu.tr

Aylin C. Küntay
Department of Psychology
Koç University
Istanbul, Turkey
akuntay@ku.edu.tr

Tilbe Göksun
Department of Psychology
Koç University
Istanbul, Turkey
tgoksun@ku.edu.tr

***Abstract*—It has been suggested that some individuals may benefit more from social robots than do others. Using second language (L2) as an example, the present study examined how individual differences in attitudes toward robots and personality traits may be related to learning outcomes. Preliminary results with 24 Turkish-speaking adults suggest that negative attitudes toward robots, more specifically thoughts and anxiety about the negative social impact that robots may have on the society, predicted how well adults learned L2 words from a social robot. The possible implications of the findings as well as future directions are also discussed.**

Keywords—human-robot interaction, second language, individual difference, robot-assisted language learning (RALL)

I. INTRODUCTION

Individual difference has been a hot topic in psychology for the past few decades [1]. Although traditional psychological research tends to focus on how humans generally think and behave, recent research has demonstrated the need for examining each individual because humans approach the same cognitive task in vastly different ways (e.g., [2]). Second language (L2) learning is no exception, and individual differences in various factors such as preference, attitudes, and personality must be considered. For example, some individuals may prefer to learn L2 through conversation with native speakers of the language whereas some others may prefer to sit alone at a desk and learn from books. Investigation of ways in which individual differences affect the process and outcomes of L2 learning is not only scientifically interesting, but also provides practical insights into how L2 learning experience can be improved by tailoring lessons for each individual learner. The current study uses robot-assisted L2 learning as an example to evaluate how individual differences predict the process and outcomes of learning, and discusses the possibility of technology facilitating learning by providing personalized lessons.

The use of social robots in education is becoming more and more popular due to improvements in their quality and affordability. Although no previous research focused specifically on the effects of individual differences in robot-assisted L2 learning, the idea has been suggested. For instance, examining word learning in fifth and sixth graders, Kanda, Hirano, Eaton, and Ishiguro (2004) found that children with some English proficiency or interest in English benefitted more from extra learning opportunities provided by social robots than did their peers with lower proficiency or interest [3]. Robots may be especially helpful for individuals with impaired social and communicative skills such as children with autism spectrum disorder (ASD). Social interactions with humans can sometimes be difficult or stressful for children with ASD because humans behave in very complex and unpredictable ways. Some researchers claim that robots can be good communication partners for those children as they can provide simpler and less stressful environments [4].

Some studies examined the relation between individual differences and how a person interacts with a robot. Ivaldi, Lefort, Peters, Chetouani, Provasi, and Zibetti (2017) examined the patterns of speech and eye gaze in 56 adults while they built an object with the humanoid robot iCub [5]. The study found that individuals who are high on extroversion tend to talk more with the robot, and individuals with a negative attitude towards robots tend to look less at the robot's face and more at the robot's hands. Tapus, Ţăpuş, and Matarić (2008) found that participants who were high on introversion interacted more with an introverted robot than an extroverted robot whereas participants who were high on extroversion interacted more with an extraverted robot than an introverted robot [6]. Takayama and Pantofaru (2009) found that having

This research was supported in part by the EC H2020 L2TOR project (grant 688014).

the personality trait of agreeableness decreases personal spaces when individuals approach robots, while having the personality trait of neuroticism and negative attitudes toward robots increase personal spaces when robots approach people [7]. These studies demonstrated that individual differences in negative attitudes toward robots and personality characteristics may predict how humans behave when they interact with a robot. However, the results are far from consistent, and more importantly, no study has examined whether individuals with different attitudes towards robots and with different personality traits learn to different levels from social robots.. Observing differences in human behaviors has scientific impact, but perhaps more important for human-robot interaction (HRI) research in individual differences is to move a step further and evaluate whether individuals with certain traits benefit more from robot companions than others. Robot-assisted L2 learning is a perfect example to explore the issue as the learning outcomes such as test scores can be directly used to evaluate how effective and beneficial the robot companion is.

To examine ways in which individual differences affect how well humans learn from or with social robots, the present study examines language learning. This article focuses on part of a larger study and reports learning outcomes of an L2 lesson and its relationship with attitudes toward robots and personality traits. We chose attitudes toward robots as a possible predictor because of the previous findings (e.g., [5,7]), and because to assess the unique nature of robot-assisted L2 lessons, it is critical to specifically examine individuals' attitudes toward robots. By assessing both negative attitudes toward robots and general personality traits such as *openness to experience* and *extroversion*, we are able to understand whether the observed relations between individual differences and learning outcomes are likely to be specific to robot-assisted L2 lessons as opposed to L2 lessons in general. For example, open-minded individuals may be more likely to learn from the robot because they are willing to interact with an unfamiliar agent and welcome the use of new technology or methods. Another possibility is that extraverted individuals benefit from any language lessons, whether with another person or a robot, because they enjoy communicating with another agent. In this study, we specifically tested the possibility that the learning outcomes of robot-assisted L2 lessons can be explained by the person's attitude towards robots. In other words, we tested the hypothesis that individuals who have positive attitudes toward robots are more likely to learn language from the lesson provided by a robot. We also tested the hypothesis that the relation is specific to attitudes towards robots and thus general personality traits such as openness to experience and extraversion do not predict the learning outcomes in robot-assisted L2 lessons.

II. PARTICIPANTS

Twenty-four Turkish speakers (*Age range* = 18.41-24.73 years; M_{age} = 20.18 years; *SD* = 1.56; 16 females) participated in the study. All participants were undergraduate and graduate students at Koç University in Istanbul, Turkey, who received course credits or monetary compensation for their participation. Participants had no known vision or hearing impairments. They were given the options of receiving monetary compensation or course credits for their participation.

III. STIMULI

A. Pre-Lesson Test and Questionnaire

Prior to the one-on-one English lesson with the NAO robot, participants completed one English test on paper and one survey on a desktop computer.

English Test. Oxford Quick Placement Test [8] was used to assess the English skills of participants. There were 60 multiple-choice questions in total.

Individual Difference Questionnaire. Total of 157 questions were prepared and all were put on one Qualtrics program to be completed on a desktop computer in the lab. This article specifically reports data from the following two sections concerning attitudes towards robots and personality traits.

- *Attitudes toward robots*. Negative Attitudes toward Robots Scale (NARS) was used to assess how participants feel about robots [9]. The NARS consists of 14 questions that can be divided into three subscales: negative attitude toward interacting with robots (S1, Questions 4, 7, 8, 9, 10, and 12), negative attitude toward social influence of robots (S2; Questions 1, 2, 11, 13, and 14), and negative attitude toward emotions involved in interaction with robots (S3; Questions 3, 5, and 6). Table I shows the Turkish version of the NARS that was developed by the first and second authors based on both the Japanese version [10] and the English version [9]. Participants rated how well each statement represents their negative attitudes toward robots on a scale of 1-5.

- *Personality traits.* Personality traits were measured based on the five-factor model of personality or "Big Five" – openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. We adapted the Turkish version of the Big Five survey used by Demir and Kumkale (2013) [11]. There were 45 questions in the survey that can be divided into five subscales: openness to experience (Questions 1-9), neuroticism (Questions 10-18), extraversion (Questions 19-27), conscientiousness (Questions 28-36), and agreeableness (Questions 37-45). Participants

TABLE I. THE TURKISH VERSION OF THE NEGATIVE ATTITUDES TOWARD ROBOTS SCALE (NARS; NOMURA, KANDA, & SUZUKI, 2006) USED IN THE PRESENT STUDY.

1	*Eğer robotların kendi duyguları olursa kaygılı hissederim.* (I will feel anxious if robots have their own emotions.)
2	*Robotların insanlara daha çok benzemesinin insanoğlu açısından olumsuz bir sonucu olacağını düşünüyorum.* (I surmise that there will be negative consequences for humans when robots become more similar to humans.)
3	*Robotlarla etkileşime girersem kendimi rahat hissederim.* (I will feel comfortable if I interact with robots.)
4	*Robotların kullanıldığı bir iş yerinde çalıştığımı hayal ettiğimde kaygılı hissederim.* (I feel anxiety when I imagine that I may be employed or assigned to a workplace where robots are used.)
5	*Eğer robotların kendi duyguları olursa kendimi onlara yakın hissederim.* (I will feel close to robots if they have their own emotions.)
6	*Robotların duygusal davrandıklarını gördüğümde kendimi daha rahat hissederim.* (I feel more comfortable when I see robots behaving affectively.)
7	*Robotlar hakkında bir şey duyduğumda bile kendimi çaresiz hissediyorum.* (I feel helpless even by hearing something about robots.)
8	*Başkalarının önünde robot kullanacak olursam kendimi utandırabilirim.* (I am likely to be embarrassed when I use robots in public.)
9	*"Yapay zekanın verdiği kararlar" veya "robotların verdiği kararlar" gibi ifadeler beni rahatsız ediyor.* (The words "artificial intelligence" or "decision by robots" make me feel unpleasant.)
10	*Sadece robotların önünde durmak bile bende gerginlik yaratır.* (Even standing in front of robots will strain me.)
11	*Robotlara aşırı bağlı olmak gelecekte olumsuzluğa sebep olabilir.* (I surmise that becoming extremely dependent on robots will have negative consequences for humans in the future.)
12	*Robotlarla etkileşime girersem kendimi tedirgin hissederim.* (I will feel nervous if I interact with robots.)
13	*Robotların çocukların zihnini olumsuz yönde etkileyeceklerinden korkuyorum.* (I am afraid that robots may negatively influence children's minds.)
14	*Gelecekteki toplumlara robotların hükmedeceği kanısındayım.* (I surmise that robots may dominate future societies.)

Note. English translations of the questions are in parentheses.

rated how well each of the statements represent their personality on a scale of 1-5.

B. English Lesson with the NAO Robot

Participants were taught eight English words – upholstery, barb, angler, caster, dromedary, cairn, derrick, and cupola. The words were selected from the last 40 items of the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4), which are supposed to be advanced for native English speakers [12]. The eight words were carefully selected so that (1) the Turkish equivalents of the words were not phonetically similar to them and (2) pronouncing the words should not be too difficult for Turkish speakers. With regard to the voice of the robot, instead of using the default Turkish text-to-speech (TTS) library in NAO, we used the female voice available on Amazon Polly ("Filiz" for Turkish and "Salli" for American English). All speech was pre-recorded as WAV sound files.

TABLE II. THE TARGET WORDS AND THEIR DEFINTIONS USED IN THE STUDY

Target word	Definition
upholstery	*Bu kelime döşemelik kumaş anlamına gelir* (This word means fabric that 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)

C. Post-Lesson Tests

Two post-lesson tests, the productive vocabulary test (hereafter the productive test) and receptive vocabulary test (hereafter the receptive test), were administered immediately after the lesson and one week later. The definitions of the target words used in the productive test were the same as the definitions used in the lesson. In the receptive test, the pictures from the PPVT that correspond to the target words were used (see Procedure for the detail of the productive and receptive tests).

IV. DESIGN

Participants were invited to the lab twice. The first visit was for the pre-lesson tests and survey (English Test and Individual Difference Questionnaire), the robot-assisted English lesson, and the immediate post-lesson tests (productive and receptive). The second visit was for the delayed post-lesson tests (productive and receptive) and the post-lesson questionnaire. The robot was controlled through a Wizard-of-Oz interface. We set one microphone behind the participant and four cameras at the corners of the ceiling, with which the "wizard" in another room monitored the participant in another room.

V. PROCEDURE

On the first visit, the participant was first asked to take the English test. Participants were given 30 minutes to complete the test although they were allowed to finish it earlier and move on to the next task. Then, participants filled out the Individual Difference Questionnaire on a desktop computer. Participants were allowed to take as much time as they needed, and it took approximately 30 minutes to complete the entire questionnaire.

After completing the English Test and the questionnaire, the participant was instructed to go into a living room-like room by herself and to sit in front of the robot. The lesson began when the NAO robot recognized the participant saying "*Merhaba* (Hello)" (Fig. 1). The robot first briefly explained the structure of lesson, and then introduced the word one by one. Each target word was taught in four steps:

1. The robot introduced the target English word and asked the participant whether she already knew the word (Note that none of the participants knew any of the target words).
2. The robot introduced the definition of the target word (see Table II).
3. The robot asked the participant to utter the target word following the robot, for three times.
4. The robot again defined the word and asked the participant to repeat the definition.

After learning every two target words, the participant was also given a mini quiz in which the robot provided the definitions of the target words and asked the participant for the corresponding word. The lesson lasted for about 20 minutes.

At the end of the lesson, the robot asked the participant to leave the room and find the experimenter who was waiting in another room. The experimenter administered the immediate productive and receptive tests. In the *productive test*, the experimenter one by one provided the definitions of the learned English words as they were defined in the lessons, 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 point to a picture that matches with the word. Participants also completed a short post-lesson questionnaire that assessed how participants felt about NAO and robots in general after finishing the lesson.

All participants were re-invited to the lab one week later (Due to schedule conflicts, the second visit took place six days after the lesson for three participants and eight days after the lesson for one participant. Due to technical issues, another participant was invited to the lab three times– once for the pre-lesson test and questionnaire, once for the lesson, and once for the post-lesson tests and questionnaire but the delay between the lesson and the post-tests was seven days). On the second visit, participants again completed the productive and receptive tests. They also completed the short post-lesson questionnaire.

Fig. 1. The participant was instructed to go into a living room-like room by herself and to sit in front of the robot. The lesson began when the NAO robot recognized the participant saying "*Merhaba* (Hello)."

VI. RESULTS

We examined whether individual differences in attitudes toward robots and personality traits predict the learning outcomes. The learning outcomes were measured in four post-lesson tests: immediate productive test, immediate receptive test, delayed productive test, and delayed receptive test. Table III shows the descriptive statistics of all analyzed variables.

First, we examined whether the scores from each of the four post-lesson tests were correlated with the ratings from each of the three subscales of NARS (S1, S2, and S3) and overall scores or the ratings from each of the five subscales (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism). As shown in Table IV, several correlations were found between the test scores and the NARS scores. Namely, the overall NARS scores were correlated with the immediate receptive test, delayed productive test, and delayed receptive test. The scores for the second subscale S2 (i.e., negative attitude toward social influence of robots) were correlated with the scores from all four tests. Among the five personality traits, openness to experience was the only one with any significant correlation, and it was only with the delayed receptive test.

Second, we built regression models to evaluate interaction among different predictors. Prior to building a regression model, correlations among possible independent variables were calculated (Table V). Significant correlations were found between (1) S1 and S2, (2) S3 and openness to experience, (3) openness to experience and extraversion, and (4) extraversion and agreeableness. To avoid the issue of multicollinearity, S1 and S3 of the NARS were excluded from the analysis. With regard to the personality traits separate models were build. For each of the four tests, three regression models were built: Model 1 included as the second scale (S2) of the NARS as a sole predictor; Model 2 included S2 as well as openness to

TABLE III. DESCRIPTIVE STATISTICS OF THE TEST SCORES AND INDIVIDUL DIFFERENCE MEASURES

	Mean	Median	SD	Min	Max
Post-Lesson Tests					
Immediate Productive	2.92	3.00	1.82	0	7
Immediate Receptive	5.13	5.00	2.03	1	8
Delayed Productive	1.46	1.00	1.59	0	5
Delayed Receptive	5.17	5.00	1.63	3	8
NARS					
S1	11.83	12.00	4.77	6	27
S2	14.92	15.50	4.42	5	23
S3	7.92	7.50	2.81	3	14
Personality					
Openess	35.79	37.00	5.56	22	44
Nauroticism	33.04	34.00	6.50	19	46
Extraversion	32.33	31.50	5.78	24	45
Conscientiousness	24.71	24.50	4.81	16	34
Agreeableness	33.71	35.00	4.78	23	45

TABLE IV. CORRELATIONS BETWEEN THE TEST SCORES AND INDIVIDUAL DIFFERENCE MEASURES

		Immediate Productive	Immediate Receptive	Delayed Productive	Delayed Receptive
NARS	All	-.35	-.45*	-.63*	-.51*
	S1	-.26	-.25	-.26	-.30
	S2	-.49*	-.53*	-.49*	-.53*
	S3	.05	-.27	.05	-.38
Personality	O[a]	.10	-.31	.10	-.44*
	N[b]	.13	-.07	.13	-.15
	E[c]	.24	-.19	.24	-.21
	C[d]	.14	-.27	.14	-.29
	A[e]	-.18	.05	-.18	.03

[a]O = Openness to experience; [b]N = Neuroticism; [e]E = Extraversion; [d]C = Conscientiousness; [e]A = Agreeableness *$p < .05$

TABLE V. CORRELATIONS AMONG THE INDEPENDENT VARIABLES

	S1	S2	S3	O	N	E	C	A
S1		.65*	.27	.06	-.18	-.02	.17	-.05
S2			.25	.11	-.15	.15	.06	-.14
S3				.46*	-.25	.40	.08	.00
O[a]					-.23	.49*	.08	-.04
N[b]						-.10	.09	-.07
E[c]							.00	-.51*
C[d]								.07
A[e]								

[a]O = Openness to experience; [b]N = Neuroticism; [e]E = Extraversion; [d]C = Conscientiousness; [e]A = Agreeableness *$p < .05$

TABLE I. TABLE VI. REGRESSION MODELS FOR IMMEDIATE PRODUCTIVE TEST

	B	*SE*	*β*	*t*
Model 1[a]				
NARS (S2)	-.24	.08	-.53	-2.95*
Model 2[b]				
NARS (S2)	-.24	.08	-.52	-2.86*
Openness	-.10	.07	-.28	-1.54
Neuroticism	-.06	.06	-.20	-1.06
Conscientiousness	-.08	.08	-.20	-1.10
Agreeableness	-.01	.08	-.03	-.19
Model 3[c]				
NARS (S2)	-.22	.08	-.54	-2.92*
Neuroticism	.02	.05	.07	.37
Extraversion	.10	.06	.33	1.79
Conscientiousness	.06	.07	.16	.88

$N = 24$; [a]Overall $R^2 = .28$; [b]Overall $R^2 = .43$; [c]Overall $R^2 = .38$; *$p < .05$.

TABLE II. TABLE VII. REGRESSION MODELS FOR THE IMMEDIATE RECEPTIVE TEST

	B	*SE*	*β*	*t*
Model 1[a]				
NARS (S2)	-.25	.05	-.71	-4.68*
Model 2[b]				
NARS (S2)	-.25	.06	-.70	-4.23*
Openness	-.04	.05	-.14	-.81
Neuroticism	.00	.04	.02	.11
Conscientiousness	.00	.05	.01	.04
Agreeableness	-.04	.05	-.11	-.67
Model 3[c]				
NARS (S2)	-.24	.09	-.52	-2.81*
Neuroticism	-.04	.06	-.14	-.75
Extraversion	-.04	.06	-.13	-.68
Conscientiousness	-.10	.08	-.23	-1.26

$N = 24$; [a]Overall $R^2 = .28$; [b]Overall $R^2 = .43$; [c]Overall $R^2 = .37$; *$p < .05$.

experience, neuroticism, conscientiousness, and agreeableness as predictors. Model 3 was built to test extraversion which could not be tested in Model 2. In addition to S2 and extraversion, neuroticism and conscientiousness were included in Model 3 as they were not correlated with extraversion.

Table VI shows the details of the regression models. According to the R^2 value, when S2 of the NARS was the sole predictor, the model explained 28% of the variance in the immediate productive test (Model 1). When openness to experience, neuroticism, conscientiousness, and agreeableness were included, the model explained 43% of the variance in the test scores (Model 2). When neuroticism, extraversion, and conscientiousness were included, the model explained 38% of the variance in the immediate productive test (Model 3). However, in all three models, S2 was the only significant predictor. The pattern was largely the same for the immediate receptive test (Table VII). In all three models, S2 was again the only significant predictor. The percentage of the variance

in the immediate receptive test explained by Models 1, 2, and 3 was 28%, 43%, and 37%, respectively.

S2 was a significant predictor for all six models built for the delayed tests. R^2 were .50, .53, and .50 for Models 1-3 of the delayed production test, and .30, .52, and .38 for Models 1-3 of the delayed receptive test. Thus, S2 explained a larger variance in the delayed tests than in the immediate tests. In addition, openness to experience was a significant predictor in Model 2 of the delayed receptive test (B = -.17; SE = .06, β = -.47; t = -2.79), suggesting that individuals with the personality trait of openness to experience tend to score low in the test.

VII. Discussion

The present study examined whether and how individual differences in attitude towards robots as well as personality traits affect learning outcomes of robot-assisted L2 lessons. We hypothesized that (1) individuals who have positive attitudes toward robots are more likely to learn L2 words from the lesson provided by a robot, and (2) the relation would be specific to attitudes towards robots and thus general personality traits such as openness to experience do not predict learning outcomes. Our preliminary data suggest that the responses to S2 of the NARS was negatively correlated with the scores of all post-lesson tests. When S2 was put into regression models with personality trait factors, S2 remained as the only significant predictor except that openness to experience was also a significant predictor in the model for the delayed receptive test. As negative attitude towards robots but not general personality traits predicted the learning outcomes, it is safe to suggest that how people learn L2 in robot-assisted lessons is affected by their attitudes toward robots.

Importantly, S2 is a scale for negative attitude toward social influence of robots, and is composed of four statements including "1. I feel anxiety if robots really have their own emotions," "2. I surmise that something negative for humans happen when robots become more similar to humans," "11. I surmise that extreme dependence on robots may cause something negative for humans in future," "13. I am afraid that robots may negatively influence children's mind," and "14. I surmise that future societies may be dominated by robots." Therefore, our results suggest that those who are afraid of robots becoming like humans and influencing human life are less likely to learn language from robots. Whereas other two scales concern participants' expectations about personal interaction with robots they themselves may experience, S2 concerns abstract fear and anxiety people have towards robots.

Although this study demonstrated the relation between learning outcomes and general and somewhat abstract negative attitudes toward robots, the mechanism underlies this relation is still unknown. We speculate that, an individual with negative attitudes toward robots is unlikely to pay attention to the robot tutor and learn well. The current data do not allow us to evaluate this possibility, and more experiments are needed to understand the relation. It is also critical to conduct the current study with human-led lessons in order to assess whether observed relations are truly specific to robot tutors. Our team is working on experiments to assess these issues in addition to recruiting more participants to the current study.

VIII. Conclusion

Researchers and educators have long been aware of the importance of recognizing individual differences. However, the topic has not received enough attention perhaps because it is unrealistic for teachers to provide personalized lessons for each individual student. Research on human-robot interaction can shed a light to the situation. By attitudes toward robots and personality traits, our study provides novel and unique insights on how robots can be used in humans learn a new language.

References

[1] Mayer, R. E. (2003). What causes individual differences in cognitive performance? In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 263-273). New York, NY, USA: Cambridge University Press.

[2] van Someren, M.W., Barnard, R., & Sandberg. J. (1994). *The think aloud method: a practical guide to modelling cognitive processes*, London, UK: Academic Press.

[3] Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, *19*(1), 61-84.

[4] Robins, B., Dautenhahn, K., & Boekhorst, R. T. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society*, *4*(2), 105-120.

[5] Ivaldi, S., Lefort, S., Peters, J., Chetouani, M., Provasi, J., & 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. *International Journal of Social Robotics*, *9*(1), 63-86.

[6] Tapus, A., Ţăpuş, C., & Matarić, M. J. (2008). User-Robot Personality Matching and Robot Behavior Adaptation for Post-Stroke Rehabilitation Therapy, *Intelligent Service Robotics Journal*, *Special Issue on Multidisciplinary Collaboration for Socially Assistive Robotics*, *1*(2), 169-183.

[7] Takayama, L., & Pantofaru, C. (2009). Influences on proxemic behaviors in human-robot interaction. *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*,

[8] G. Syndicate, U. C. L. E. (2001). *Quick Placement Test*. Oxford, UK: Oxford University Press.

[9] Nomura, T., Kanda, T., & Suzuki, T. (2006). Experimental Investigation into Influence of Negative Attitudes toward Robots on Human-Robot Interaction. *AI & Society*, *20*(2), 138-150.

[10] Nomura, T., Kanda, T., Suzuki, T. Yamada, S., & Kato, K. (2010). Human Attitudes, Anxiety, and Behaviors in Human-Robot Interation (HRI). *Proceedings of the 26th Fuzzy System Symposum*, 554-559.

[11] Demir, B., & Kumkale, G. T. (2013). Individual differences in willingness to become an organ donor: A decision tree approach to reasoned action. *Personality and Individual Differences*, *55* (1), 63-69.

[12] Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test - Fourth edition (PPVT)*. Minneapolis, MN: NCS Pearson.

Sosyal robotların jest kullanımının çocuklarda ikinci dil öğrenimine etkileri

Koç Üniversitesi, Iowa Üniversitesi

Ö. Ece Demir-Lira, Cansu Oranç, Junko Kanero, Sümeyye Koşkulu, İdil Franko, Zeynep Adıgüzel,

Tilbe Göksun ve Aylin C. Küntay

Giriş

- Konuşmacıların ürettiği el jestlerin hem konuşmacının kendi düşüncelerini aktarması ve iletişimi hem de dinleyicinin algısı üzerinde bilişsel açıdan olumlu etkileri vardır. Fakat bu olumlu etkiler bazı durumlarda görülmemektedir.
- İletişimsel iki ana jest türü işaret ve ikonik jestleridir.
  - İşaret: işaret parmağı ile çevredeki bir nesne, yer veya kişiyi gösterir.
  - İkonik: bir hareket, şekil, yer veya kişiyi tasvir etmek için kullanılır.
- Robotların hareket becerileri onları dil eğitiminde önemli bir bilgi kaynağı durumuna getirmektedir. Buna karşın, robotların jest kullanımlarının etkilerine dair detaylı çalışmalar bulunmamaktadır.

Çalışmanın amacı

- Bir sosyal robot olan NAO kullanılarak iletişimsel jestlerin üç farklı türünün (işaret ve ikonik jestleri) çocuklarda İngilizce sözcük öğrenmeyi nasıl kolaylaştırdığına anlamak
- Robotların jestleri ile dikkat vurgulayıcı diğer öğeleri karşılaştırarak robot jestlerin öğrenmede özel bir rol mü oynadığını yoksa sadece dikkat çektiklerini mi anlamak
- Sosyal robot NAO ve insan tarafından ikinci dil öğretimini karşılaştırmak

Yöntem

Katılımcılar:

- Anadili Türkçe olan 5-6 yaşındaki çocuklar
- Her grupta 20 kişi olmak üzere toplam 80 çocuk
- Her çocuk iki eğitmen koşulundan (NAO, insan), sonra da iki jest koşulundan (ikonik, işaret) birine dahil edilecektir ve ek olarak dikkat koşulunda da yer alacaktır

Malzeme ve Ölçümler:

- Öğretilecek sözcükler: 8 İngilizce ölçüm sözcükleri
  - big-small (büyük-küçük), long-short (uzun-kısa), wide-narrow (geniş-dar), high-low (yüksek-alçak).
- Çocukların bu sözcüklerin Türkçe çevirilerine ilişkin bilgileri
- Sözel ve uzamsal kısa süreli bellek

Deney Tasarımı ve İşleyiş:

- Çocuklara NAO'nun onlara İngilizce sözcükler öğreteceği söylenecektir.
- Sözcükleri farklı nesnelerle gösterilecektir (örn., büyük top, uzun çiçek).
- Kavramları ifade eden nesnelerin resimleri NAO ile çocuk arasına yerleştirilecek bir ekranda sunulacaktır (Şekil 1a).
- Eğitmen sözcükleri üç koşulda sunacaktır: ikonik jest (IJ), işaret jest (SJ), dikkat vurgulayıcı (DV)
- IJ: Sözcüğe bir ikonik jest eşlik edecektir (örn., "big" sözcüğü iki elin üst gövdenin iki yanına açılmasıyla ifade edilecektir, Şekil 1b).
- SJ: Eğitmen ekrandaki görseli işaret edecektir
- DV: Eğitmenin elleri sabit kalacak, tablet ekranındaki nesneye kırmızı bir çerçeve ile dikkat çekilecektir (Şekil 1c).

Şekil 1

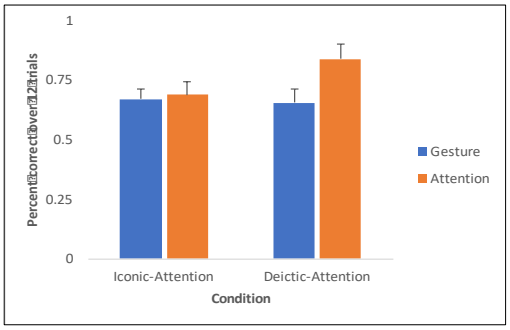
A: Büyük-küçük görseller | B: Küçük için ikonik jest | C: Büyük için dikkat vurgulayıcı

- Her sözcük çifti 3 kere tekrar edilecektir. Çocukların öğrenmeleri her tekrardan sonra öğrendikleri sözcüklere karşılık gelen görselleri seçmeleri istenerek ölçülecektir. Cevapları 1-doğru 0-yanlış şeklinde kodlanacaktır.
- Jestlerin süreleri ve sözcüklerin dilbilimsel özellikleri gruplar arası denkleştirilecektir.

Sonuçlar

İnsan eğitmen sonuçları:

- Çocukların cevapları ANOVA ile analiz edilmiştir, Jest (Ikonik, İşaret) ve Gösterme (El jesti, Dikkat vurgulayıcı)
- Çocuklar Dikkat vurgulayıcı koşulunda Jest koşuluna oranlar daha fazla doğru cevap vermişlerdir ($F(1,33) = 4.11$, $p = .05$).
- Jest ve Gösterme koşulları arasında bir etkileşim görülmüştür, ($F(1,33) = 2.69$, $p = .11$).
  - Dikkat vurgulayıcı koşulu bu koşul gösterme jestlerini takip ederse daha iyi performansa yol açmıştır.
- Jest koşulun bir etkisi görülmemiştir ($F(1,33) = 1.04$, $p = .32$).



NAO eğitmen sonuçları:

- Data toplamı devam etmektedir
- Planlanan analizler
  - Jest ve Gösterme koşullarının performans üzerindeki etkileri incelenecektir.
  - Performans üzerinde çocukların yaşının, sözel ve uzamsal belleğinin etkisi ölçülecektir.
- Beklenen sonuçlar
  - NAO'nun jestleri çocukların ilgisini daha çok çektiği için daha iyi performansa yol açabilir
  - Öte yandan NAO jestlerini ilk kez gören çocuklar bu jestlerden bir insandan ögrendikleri kadar öğrenemeyebilirler.

Tartışma

- Çalışmanın bulguları jestlerin neden öğrenmeyi kolaylaştırdığına dair bilgi sağlayacaktır.
- Jest koşulları arasında anlamlı bir fark bulunmamalıdır.
- Kolaylaştırıcı etkinin jestlerin dikkat çekmesinden kaynaklanması durumunda ise jest koşulları dikkat koşulundan farklı bir etki
- Bu çalışma robot jestlerin öğrenme üzerindeki etkilerine dair önemli teorik katkıda bulunmaktadır.
- Böylece, robotların çocuklara ikinci dil öğretmelerine ilişkin gelecekte yapılacak çalışmalara zemin oluşturmaktadır.z

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 688014.

Kaynakça

1. Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. Psychological bulletin, 137(2), 297.
2. Singer, M. A., & Goldin-Meadow, S. (2005). Children learn when their teacher's gestures and speech differ. Psychological Science, 16, 85-89.
3. Tellier, Marion. "The effect of gestures on second language memorisation by young children." Gesture 8, no. 2 (2008): 219-235.
4. Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. Human brain mapping, 32(6), 982-998.
5. Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. Language and cognitive processes, 24(2), 313-334.
6. Novack, M., & Goldin-Meadow, S. (2015). Learning from gesture: how our hands change our minds. Educational psychology review, 27(3), 405-412.
7. Kanero, J., Geçkin V., Oranç, C., Mamus, E., Küntay, A. C., & Göksun, T. (2018). Social robots for early language learning. Child Development Perspectives. doi: 10.1111/cdep.12277
8. Conti, D., Di Nuovo, A., Cirasa, C., & Di Nuovo, S. (2017, March). A comparison of kindergarten storytelling by human and humanoid robot with different social behavior. In Proceedings of the Companion

2 2017 Publications

- Junko Kanero, Mirjam de Haas, Ezgi Mamus, Cansu Oranç, Rianne van den Berghe, Josje Verhagen, Ora Oudgenoeg-Paz, Kirsten Bergmann, Thorsten Schodde, Aylin C. Küntay,Tilbe Göksun, Paul Vogt (2017, October). Observing human tutoring to develop robot-based language lessons. *Symposium on Multimodal Communication 2017,* Bielefeld, Germany.

- Sebastian Wallkötter, Michael Joannou, Samuel Westlake and Tony Belpaeme (2017, October). Continuous Multi-Modal Interaction Causes Human-Robot Alignment. In *Proceedings of the 5th International Conference on Human Agent Interaction (pp. 375-379),* ACM, Bielefeld, Germany.

- Christopher D. Wallbridge, Séverin Lemaignan and Tony Belpaeme (2017, October). Qualitative Review of Object Recognition Techniques for Tabletop Manipulation. In *Proceedings of the 5th International Conference on Human Agent Interaction (pp. 359-363),* ACM, Bielefeld, Germany.

- Thorsten Schodde, Laura Hoffmann, and Stefan Kopp (2017) How to Manage Affective State in Child-Robot Tutoring Interactions? In *Proceedings of IEEE International Conference on Companion Technology (ICCT 2017),* Ulm, Germany.

- Séverin Lemaignan, Charlotte Edmunds, Emmanuel Senft, and Tony Belpaeme (2017). The Free-play Sandbox: a Methodology for the Evaluation of Social Robotics and a Dataset of Social Interactions. *arXiv preprint arXiv:1712.02421.*

- James Kennedy, Paul Baxter, and Tony Belpaeme (2017). Nonverbal Immediacy as a Characterisation of Social Behaviour for Human-Robot Interaction. *International Journal of Social Robotics,* 9: 109.

- Paul Baxter, Emily Ashurst, Robin Read, James Kennedy, and Tony Belpaeme (2017). Robot education peers in a situated primary school study: Personalisation promotes child learning. *PLoS One, 12(5),* e0178126.

- James Kennedy, Paul Baxter, and Tony Belpaeme (2017). The Impact of Robot Tutor Nonverbal Social Behaviour on Child Learning. *Frontiers in ICT, 4, 6.*

- Rianne Vlaar, Josje Verhagen, Ora Oudgenoeg-Paz, and Paul Leseman (2017) Comparing L2 Word Learning through a Tablet or Real Objects: What Benefits Learning Most? In *Proceedings of the Workshop R4L,* at ACM/IEEE HRI 2017.

- Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt (2017) Exploring the Effect of Gestures and Adaptive Tutoring on Childrens Comprehension of L2 Vocabularies. In *Proceedings of the Workshop R4L,* at ACM/IEEE HRI 2017, 2017.

- James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17).* ACM, New York, NY, USA, 82-90.

- Mirjam de Haas, Peta Baxter, Chiara de Jong, Emiel Krahmer, and Paul Vogt (2017) Exploring Different Types of Feedback in Preschooler and Robot Interaction. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17).* ACM, New York, NY, USA, 127-128.

- Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp (2017) Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. In *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI'17).*

- Wafa Johal, Paul Vogt, James Kennedy, Mirjam de Haas, Ana Paiva, Ginevra Castellano, Sandra Okita, Fumihide Tanaka, Tony Belpaeme, and Pierre Dillenbourg (2017). Workshop on Robots for Learning: R4L. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17).* ACM, New York, NY, USA, 423-424.

- Paul Vogt, Mirjam de Haas, Chiara de Jong, Peta Baxter and Emiel Krahmer (2017) Child-Robot Interactions for Second Language Tutoring to Preschool Children. *Frontiers Human Neuroscience* 11,73. DOI: 10.3389/fnhum.2017.00073

- Peta Baxter, Chiara de Jong, Rian Aarts, Mirjam de Haas, and Paul Vogt (2017). The effect of age on engagement in preschoolers child-robot interactions. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI '17).* ACM, New York, NY, USA

- Vasfiye Geçkin, Ezgi Mamus, Cansu Oranç, Başak Güven, Aylin C. Küntay, and Tilbe Göksun. (2017, April). Development of spatial concepts in a second language: Can feedback defeat complexity?. Poster presented at *the Biennial Meeting of the Society for Research in Child Development,* Austin, TX. *

- Vasfiye Geçkin, Ezgi Mamus, Cansu Oranç, Junko Kanero, Aylin C. Küntay, and Tilbe Göksun. (2017, August). L2 Teachers' Verbal and Nonverbal Orchestration in Preschools: Implications for Educational Robots. Poster presented at *the 18th European Conference on Developmental Psychology,* Utrecht, Netherlands. *

*No pdf available

- Paul Vogt, Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, and Emiel Krahmer (2017). Adaptation and gestures in second language tutoring using social robots. *Workshop on early literacy and (digital) media.* *

*No pdf available

Observing human tutoring to develop robot-based language lessons

Decades of research by psychologists and educationalists identified a number of strategies human adults use to teach a new language to young children (e.g., Konishi et al., 2014). In recent years, scholars in robotics and related fields have also been involved in research on early language education, advocating the potential of humanoid robots as companions that simulate the way human adults teach language (e.g., de Haas et al., 2016). To this date, however, there has not been extensive discussion on how strategies employed by human teachers can be applied to develop robot-based language lessons. In this project, we aim to determine which pedagogical approaches can be and should be implemented in robot-based language lessons.

This abstract discusses our first step – analysis of teaching strategies observed in preschool language classes, with the special focus on the use of first (L1) and second (L2) languages and bodily actions such as gestures. We chose these two topics due to the potential strengths of a robot as a language teaching tool. First, the ability to switch between L1 and L2 makes robots as effective as or perhaps more effective than human teachers in some situations. It is often difficult for a human teacher to switch between two languages especially when the classroom consists of children from different language backgrounds. A robot can provide supplementary one-on-one L2 lessons using any L1-L2 combinations. Second, the capability to perform actions makes a robot different from other devices such as a tablet. As a physical agent with arms and legs, humanoid robots are able to perform various gestures which are, at least when performed by humans, known to facilitate language learning in children (e.g., Hostetter, 2011; Sueyoshi & Hardison, 2005).

To assess how human L2 teachers use language and actions, we conducted semi-naturalistic observations of (1) large-group L2 English lessons at preschools in Turkey, (2) one-on-one or small-group L2 English lessons in the Netherlands and Germany, and (3) L2 Dutch lessons for children from immigrant families in the Netherlands.

First, teachers' utterances were transcribed, and all bodily actions were noted alongside. We then coded each utterance for a number of characteristics using an original coding scheme. All utterances were coded for whether it was in L1 or L2, and for whether switching between L1 and L2 occurred. Gestures and other actions were classified at different levels. At a global level, gestures can be classified into categories such as a *deictic gesture* (pointing at different entities such as objects or locations), an *iconic gesture* (gesture that represents a concrete event or object), or a *metaphoric gesture* (gesture that represents an abstract concept such as knowledge). It was also useful for our purpose to note more specific categories that can be directly used in our robot-based lessons. Thus, our codes included both general categories (e.g.,deictic gestures) and specific categories (e.g., pointing to a box, pretending to wear a jacket). Some of these codes were derived from the literature whereas others were added by our coders based on the observations. We also coded non-gestural actions (e.g., dancing on a song) because most commercially available humanoids (e.g., Softbank Robotics NAO) are expected be able to perform them.

Our observations show that, in terms of language use, English teachers in Turkey and the Netherlands mainly used the L2 as the medium of instruction. However, the teachers sometimes shifted from L2 to L1 (1) to manage classroom issues, (2) to ask questions, (3) to give instructions, and (4) to explain syntactic or phonological rules (e.g., explaining the difference between 'this is' vs. 'these are' or explaining 'the singular-plural distinction' as in shoe vs. shoes). However, in Germany, the teacher switched very frequently between the L1 and L2: out of all utterances: 55% was in L1, 30% was in L2, and 15% was unclassifiable (e.g., interjection, children's names). We can claim that teachers were naturally adjusting

their language in order to ensure that children understood key concepts. Although many L2 programs take, or at least claim to take, a total immersion approach in which the teacher speaks only in L2, the use of L1 can be still observed and is believed to be quite beneficial in some situations (e.g., Moore, 2002).

The use of gestures and other actions was very frequent in all lessons. The amount, however, varied greatly across lessons, from 9.24 to 73.07 per 20 minutes. Importantly, the rate of action use seemed to depend largely on the theme of a lesson. For instance, when teaching names of body parts, 70% of the teacher's utterances containing target words were accompanied with gestures (e.g., pointing to the arms), but when the lesson theme was weather, the teacher used gestures in only 18% of her target word utterances. Thus, teachers used gestures only when there was a conventional or very straightforward gesture associated with the word they were teaching.

Our data suggest that, although the mere presence of gestures may increase children's attention to the learning content (e.g., Hostetter, 2011), including gestures in every possible occasion may not be necessary. Gestures are suited to teach words in some domains such as math (Cook & Goldin-Meadow, 2006), but may not be as important when teaching concepts with no conventional gestures such as body parts or color because the behavior was not observed among human teachers either. It must be noted, however, in teaching any concepts, some gestures such as pointing can be useful. Pointing can direct children's attention to any relevant object, material, or location. In fact, pointing was more common than any other gestures (e.g., iconic gestures) in our class observation.

So how can we use the information in L2 classrooms to develop a robot L2 tutor? Our data on language use suggests that L1 is used in L2 classrooms more often than commonly believed, and the amount of L1 use is flexibly determined based on various factors such as lesson topics and L2 proficiency of students. The results also highlight the potential benefits of using a robot as a language tutor because, as mentioned earlier, a robot can be programmed to use any combination of L1 and L2 in theory.

Translation of the pedagogical strategies used by human teachers to robot-based lessons also introduce unique challenges. Although we found that the teachers constantly performed actions to facilitate their learning process, the robot gesturing too much might cause more harm than good. Most humanoids available under the status quo cannot move as flexibly or smoothly as humans, and thus some of the gestures observed in the classrooms cannot be well replicated by robot tutors. Further, many robots produce motor sounds while gesturing and thus can mask speech sound when utterance and gesture simultaneously occur. Research suggests that overuse of actions and gestures or mismatch between speech and gesture can impede the word learning process (e.g., Goldin-Meadow & Sandhofer, 1999). Even human teachers do not perform actions in some situations, and thus in designing robot-assisted L2 lessons, we must carefully consider when the use of actions and gestures is truly appropriate, as opposed to including them as much as possible.

In conclusion, we emphasize that observation of human tutoring can be quite beneficial in developing robot learning companions not only because it provides general ideas about how children learn a new language, but also because specific phrases and actions used by human teachers are most likely to be familiar for children and thus may help children recognize the robot tutor as an agent and to have a successful learning experience. With regards to some features such as gestures, we must carefully consider the balance between what we want the robot to do and what hardware and software limitations of the particular robot let us do. Observations of human tutoring can serve as a good starting point in determining what to consider in developing educational robots.

Continuous Multi-Modal Interaction Causes Human-Robot Alignment

Sebastian Wallkötter
Centre for Robotics and Neural Systems
Plymouth, United Kingdom
sebastian.wallkotter@postgrad.plymouth.ac.uk

Michael Joannou
Centre for Robotics and Neural Systems
Plymouth, United Kingdom
michael.joannou@postgrad.plymouth.ac.uk

Samuel Westlake
Centre for Robotics and Neural Systems
Plymouth, United Kingdom
samuel.westlake@postgrad.plymouth.ac.uk

Tony Belphaeme
Centre for Robotics and Neural Systems
Plymouth, United Kingdom
tony.belpaeme@plymouth.ac.uk

ABSTRACT
This study explores the effect of continuous interaction with a multi-modal robot on alignment in user dialogue. A game application of '20 Questions' was developed for a SoftBank Robotics NAO robot with supporting gestures, and a study was carried out in which subjects played a number of games. The robot's confidence of speech comprehension was logged and used to analyse the similarity between application legal dialogue and user speech. It was found that subjects significantly aligned their dialogue to the robot throughout continuous, multi-modal interaction.

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces; I.2.9. Artificial Intelligence: Robotics; I.2.6. Artificial Intelligence: Learning

Author Keywords
human-robot alignment; multi-modal interaction; SoftBank NAO Robot

INTRODUCTION
Whether interacting with a child, a colleague or a stranger, it is widely accepted that humans adapt their communication in accordance with their understanding of the listener's knowledge and capability [1]. This unconscious process occurs across multiple channels, and greatly simplifies dialogue production and comprehension [6]. In contrast, many modern domestic robots still use just a single modality which can result in the loss of information, such as context, and less effective communication and irritation for the user.

HAI '17, October 17–20, 2017, Bielefeld, Germany

DOI: https://doi.org/10.1145/3125739.3132599

With emergence of less computationally-expensive computer vision techniques and advances in the field of HRI, modern social robots can utilise some of the many non-verbal forms of communication that come naturally to humans. For example, gesture and gaze comprehension are of particular importance when resolving context in dialogue, as humans often refer to objects and events using these channels. This report explores the extent to which humans naturally align to multi-modal, human-like robot communication.

The strength of this alignment will have implications in the design of future HRI systems. In this report, a study was conducted to understand the strength of human verbal alignment, i.e. adaptation of grammar, vocabulary and speaking style, to a multi-modal social robot by continuous interaction. A SoftBank Robotics, NAO robot was programmed to play games of '20 Questions'. Users would think of an animal and the NAO would work out what the animal was by asking a series of questions. In addition to verbal communication, the robot was capable of relaying information through LEDs in its eyes and ears, and via gestures. Interactions between the user and the robot were then logged over a series of games in order to evaluate if humans automatically adapt to robots even when the robot utilises multiple communication channels.

RELATED WORK
Entertainment is one of the most promising applications of social robots [4]. However, the consequences of fragile, error-prone communication systems in HRI include degraded performance and limited commercial potential [5]. Dialogue performance can be greatly improved through the additional utilisation of non-verbal modalities, as demonstrated by a study involving a storytelling robot [4].

Another application of social-robots is their use as classroom assistants. When designing a robot to aid children in their learning, one would readily assume that the robot, like human tutors, should have social and adaptive behaviour. However, experiments by Kennedy, Baxter and Belpaeme [2] demonstrated that this is not necessarily the case, and it was hypoth-

esised that social behaviour of robots may distract children from their learning tasks.

An important factor to consider in HRI is the way in which users align their verbal communication style to the robot, particularly in error-resolution situations. Oviatt, Bernard and Levow [5] analysed the type and magnitude of linguistic adoption that occurred during human-computer error resolution. They discovered that users adapt to the system in three distinct ways: increasing linguistic contrast, increasing hyperarticulation and suppression of linguistic variability. Further, the researchers also found that the feedback given by the robot had a significant effect on the users' behaviour [3].

IMPLEMENTATION

Hardware

The NAO robot was chosen for this experiment as it has broad functionality and an infant-like appearance that helps to limit any preconceived expectations of its capabilities. It has 25 degrees of freedom to allow the development of a range of physical gestures to complement verbal communication. There are 11 degrees of freedom in the lower body (pelvis and legs) and 14 in the upper body (head and arms). Low level control is updated every millisecond while high level control and sensor data is updated every 20 *ms*. Additional features include two loudspeakers to allow the robot to play audio and speech, as well as four microphones (two at the front of the head and two at the back) to allow the robot to capture the user's speech. Captured utterances are processed using NAO's built-in speech recognition engine (Nuance VoCon 4.7).

Software

Gestures

Blinking is a subconscious form of non-verbal communication in human-human interaction and consequently, prolonged staring throughout an interaction will result in alienation. The NAO has a number of LEDs embedded in its head, with a total of 24 LEDs dedicated to each eye. To maximise agency, LEDs were turned off and on in sequence to imitate human blinking which occurred at a constant base frequency with added random noise. Throughout normal interaction, the NAO's eyes shone white; however, upon comprehending speech with high confidence, the eyes flashed green for one second. Conversely, if detected speech was not confidently comprehended, the eyes flashed red. This modality was designed to play a major role in informing the user how to adapt for alignment. Additionally, LEDs in the NAO's ears were turned on upon detection of sounds above a certain volume threshold and otherwise, turned off. Given that verbal feedback can be invasive to conversation, these LED controls were implemented to provide an intuitive alternative.

Further gestures were implemented by manipulating the robot's joints. Upon receiving an answer from the user, a motion to suggest that the NAO was thinking was selected at random, initiated and coupled with a verbal response. Question-specific gestures were also implemented as well as end-game gestures that represented the NAO's reaction to either losing or winning the game. The advantages of this approach were twofold. Firstly, this approach forced breaks in the conversation and gave the dialogue a more natural pace, closer to that of human-human interactions. Secondly, these motions gave the user some indication of what the robot is doing, namely, processing the answer of the previous question, and indicated that the robot will give a response in a moment.

QiChat

The corpus was outlined within QiChat topic files using a bootstrap method. The resulting system was context-based grammar, and consequently, only a restricted portion of the grammar was available at any particular point depending on the flow of the conversation. This was achieved by dynamically loading and unloading portions of the corpus.

The overall dialogue flow was system initiative but could be switched to short user initiative dialogues upon particular user requests. To encourage the user to stay within the grammar it was decided that in-corpus grammar would be used when the robot was talking. Possible questions the NAO could ask were specified in YAML files along with the grammar for the expected answers. Once the user response to the question was received and recognised, it was passed to a Python script running the game engine.

Each question topic contained a concept for 'yes' and 'no' that allowed the question to be answered in a variety of ways specific to that question. Additional topics were added to handle uncertainty in sentences (e.g. 'I think so'). This was to ensure that such an answer does not result in the disqualification of a possible animal due to gaps in the user's animal knowledge or wrong answers. For instance: NAO: 'Does it fly?' User: 'Maybe'. This would not disqualify the animal 'bird' from the list of potential candidates.

Although animals are not necessarily disqualified due to user responses, the nature of the animal YAML file definitions ensured that some uncertainty is accounted for. Each animal definition was outlined in its own YAML file. The file contained the name of the animal, a short question to be asked when the robot wished to guess the animal (e.g. 'is it a bear?') and a frequency value for each label of each question the robot may ask. This frequency represents the number of times a label, animal pair has been observed in the past. For instance a cat may have a frequency value of '100' for 'it has fur' and a value of '5' for 'it does not have fur'. This adds robustness to the system given an instance when the user says 'no' as the cat is not completely disqualified.

Game Engine

This section describes the robot's internal representation of the game. First, answers to questions were clustered into a finite amount of categories called labels, L. A question such as 'Does it fly?' would have two: 'yes' and 'no'. Elements of the robot's corpus were then be mapped onto the according label. Questions were subdivided into two categories: differentiating questions and guesses. Differentiating questions, Q, as the name implies, help the robot to differentiate between animals. An example would be: 'Does it have legs?'. Guesses G, are yes/no questions, specifically asking for an animal, e.g. 'Is it a cat?'. To win the game, the robot has to ask a guess and detect

the label "yes". There was a total of $Q = 17$ differentiating questions and $G = 32$ guesses, one for each animal, making a total of 49 questions. All question labels were combined into a feature space, where each dimension represented the frequency of a label's observation. Each animal was then represented as an element of this space.

To find the animal, it was assumed that the user's animal was in the set of animals, A, known to the robot. This allowed the robot to create a probability distribution over animals, modelling the user's belief. The optimal distribution, P^* would assign 0 probability to every animal except the user's, which would have probability 1. This distribution had to be found by asking questions. Initially however, the robot had no information about the user's animal, thus its prior, $P(A)$, was a uniform distribution over all animals. Given the label of the user's response to a question, this prior could be improved in a Bayesian fashion:

$$P(A|L=l) = \frac{P(L=l|A)P(A)}{P(L=l)}\ , \tag{1}$$

where $P(L=l)$ was the total probability of observing label l as an answer and $P(L=l|A)$ was the probability of observing l, given the currently asked question. Further, $P(A)$ was the robot's current prior and $P(A|L=l)$ was the new, better prior. As an alternative to computing $P(L=l)$ and marginalising over it, one can normalise the result of $P(L=l|A)P(A)$ after computation.

One challenge in using this method alone was that the probability of an animal could never reach 0 exactly. This decreased robustness if the robot's model of an animal differed from the user's. An example would be the user thinking of a squirrel and being asked: 'Does it have two or four legs?'. While the robot may think that a squirrel has two legs, the user may think it has four and thus answer accordingly. This would decrease the squirrel's probability and other animals, i.e. dog, cat, and so on would become more likely. To solve this, the robot's belief was thresholded after each Bayesian update:

$$P_{\text{thresh}}(A) = \begin{cases} P(A)\ ,\ \text{if } P(A) \geq \frac{0.05}{\sum_A [P(A)>0]} \\ 0\ ,\ \text{otherwise} \end{cases}\ . \tag{2}$$

Here, $\sum_A [P(A) > 0]$ counted the number of animals with a non-zero likelihood, scaling the threshold dynamically. This can be viewed as 'discarding' an animal, if enough information had been gathered suggesting another. Not only did this solve the problem of a potential difference between the robot's and the user's model of an animal; tests also showed that this creates robustness against deliberately-supplied false information. For example, if the only remaining animals are a cat or a dog, both of which are equally unlikely to fly, then if the user told the robot that the animal does fly, each animal's probability would decrease initially, but reset after normalisation. This means the robot is mostly unaffected by false information, if enough correct information has been specified beforehand.

Finally, the robot chose its next question depending on how many animals could be discarded on average when asking. It was done by simulating each label as a reply for each question, using above inference method. However, when computing the updated prior, P_{thresh}, the number of times a probability was set to 0 was counted. As the likelihood for a label is known, the expected number of discarded animals per question could be calculated. Consequentially, the question with the highest expected value was chosen as next question.

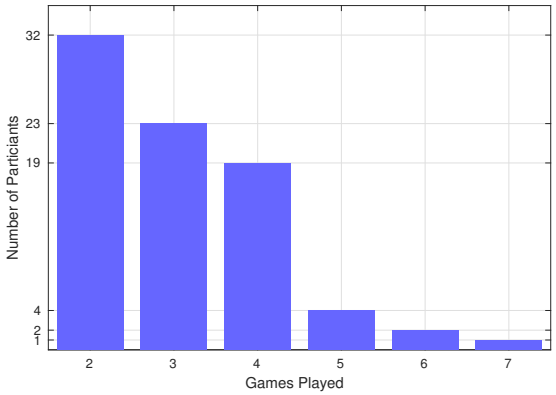


Figure 1. The figure shows the distribution of games played by participants. There was a total of 32 participants. The graph shows how many participants played at least N games.

This setup scales well into the case of an unknown animal. The robot would assign high probability to a known animal sharing the most features with with user's animal. However, since the user will answer the corresponding guess question with label 'no', the robot runs out of animals to consider and concedes.

STUDY DESIGN

The goal of this paper is to answer the hypothesis: 'Does continuous multi-modal interaction cause human-robot alignment?'. A within-subject study was conducted, asking a number of subjects to play a sequence of four games. For each game the subject's verbal alignment to the robot was measured.

In the beginning, the robot would offer an explanation of the game and then start the experiment. This allowed a controlled and repeatable introduction. During the experiment the robot would record all detections of the speech recognition engine together with their confidence. This capture happened automatically and in the background, minimising influence on the subject.

The way the study was set up allowed minimal interaction between the researchers and the subject. This provided consistency across all experiments and minimised the Hawthorne effect as neither observers nor clear recording equipment (camera or microphone) were present. This facilitated authentic or near authentic behaviour throughout the interaction.

RESULTS AND ANALYSIS

Carrying out the study, a total of 32 subjects were asked to play initially. However, many participants could not play four subsequent games, due to time constraints. The distribution over how many games were played by all participants is shown in figure 1. Each of the 32 participants played at least two games,

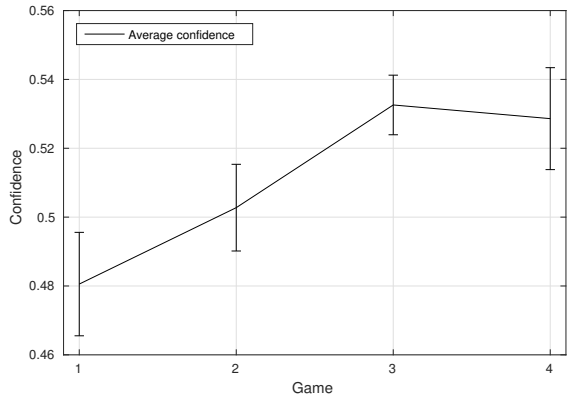


Figure 2. The graph shows the average confidence of the speech recognition system over the number of games played. The error bars visualize the standard error. The confidence increases significantly ($p = 0.018$) over the course of multiple games.

however only a total of 19 played four or more consecutive games.

The group of participants that played four or more games was analysed using ANOVA with repeated measurements. For this, the first four games of each participant were considered. Figure 2 shows the average confidence in each game as well as the standard error. The result shows that the confidence increases significantly over time ($F(3,54) = 3.651$, $p = 0.018$).

To measure the speech recognition's confidence the ASR's (Nuance VoCon 4.7) confidence value was used. This result suggests alignment between the human and the robot.

DISCUSSION

Throughout these interactions, the principal method of indicating if detected speech had been matched to a phrase in the corpus was non-verbal and expressed via the colour of the robot's eyes. In addition, only implicit verbal feedback was given by the NAO as to whether the subject's answer was correctly categorised by the dialogue system. This ensured that no information was given as to the specific content of the robot's grammar. Users acquired knowledge of the NAO's grammar through trial and error only, and therefore, alignment occurred entirely naturally, without explicit instruction from the robot or a researcher.

As seen in Figure 2, the NAO's confidence of comprehension initially averaged at a level below the speech recognition confidence threshold of 0.5. As interaction continued, the average confidence of comprehension increased significantly, and eventually peaked in game number three. This experiment demonstrated that subjects significantly aligned their spoken communication during these multi-modal interactions to maximise the NAO's confidence of comprehension.

A slight, but insignificant, decline in the confidence of comprehension was observed in game number four, seen in Figure 2. The reason behind this is unknown. However, the answer may lie in an underlying compromise between effective dialogue, and speech that is natural to the user. Significant alignment occurred throughout the first three games, at which point the conversation may be considered effective, however, it is likely that subjects suppressed their natural linguistic variability to achieve this. Once sufficiently effective dialogue had been achieved, the users may have begun to slip back into more natural linguistic habits.

The results of this study should be leveraged by designers of social robots. The strong degree of alignment that was observed indicates that subjects quickly built a belief of the robot's capability in order to predict what the robot will understand and, subsequently, tailor their grammar accordingly. Consequently, this implies that small corpora can still result in efficient dialogue, whilst reducing development time. The occurrence of significant alignment implies that the NAO was 'over-promised', a situation that can lead to disappointment for the user. Consequently, this report hypothesises that gradient of alignment can be used as proxy for measuring the degree to which a robot has been over-promised.

CONCLUSION

This study found that subjects automatically strayed from their natural style of verbal communication in order to align their dialogue with that of the NAO robot throughout continuous, multi-modal interaction. This adaptation occurred in the presence of communication through multiple channels, with the NAO relaying information through speech, gestures and LEDs in its eyes and ears. In addition, this alignment occurred in the absence of explicit instruction from the robot or researchers.

The study observed some degree of overshoot when subjects simplified their speech to align with the robot. However, this was not observed with statistical significance. If true, it would highlight the compromise that users make between effective interaction and natural speech.

It is clear that the phenomenon of alignment has positive and significant effects on the effectiveness of dialogue in HRI. This paper proposes that the gradient of alignment could also be used as a proxy to measure the degree to which a robot is over-promised by its appearance. Future study into the possible interaction between rate of alignment and over-promising is recommended.

REFERENCES

1. Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9 (2010), 2355–2368.
2. James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. (2015), 67–74.
3. Manja Lohse, Katharina J Rohlfing, Britta Wrede, and Gerhard Sagerer. 2008. "Try something else!"–When users change their discursive behavior in human-robot interaction. (2008), 3481–3486.
4. Bilge Mutlu, Jodi Forlizzi, and Jessica Hodgins. 2006. A storytelling robot: Modeling and evaluation of human-like gaze behavior. (2006), 518–523.

5. S Oviatt, J Bernard, and G Levow. 1998. Linguistic adaptations during spoken and multimodal error resolution. *Language and speech* 41, 3-4 (1998), 419–442.
6. Martin J Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences* 27, 2 (2004), 169–190.

Qualitative Review of Object Recognition Techniques for Tabletop Manipulation

Christopher D. Wallbridge, Séverin Lemaignan and Tony Belpaeme
Plymouth University
Plymouth, UK
{christopher.wallbridge, severin.lemaignan, tony.belpaeme}@plymouth.ac.uk

ABSTRACT

This paper provides a qualitative review of different object recognition techniques relevant for near-proximity Human-Robot Interaction. These techniques are divided into three categories: 2D correspondence, 3D correspondence and non-vision based methods. For each technique an implementation is chosen that is representative of the existing technology to provide a broad review to assist in selecting an appropriate method for tabletop object recognition manipulation. For each of these techniques we give their strengths and weaknesses based on defined criteria. We then discuss and provide recommendations for each of them.

ACM Classification Keywords

I.4.8 Scene Analysis: Object Recognition

Author Keywords

object detection; pose detection; tabletop manipulation.

INTRODUCTION

Context: Near Object Interaction

This paper takes a practical approach to survey the technical landscape on the problem of small object identification and 6D object localisation in a cluttered environment – a context often termed as *object recognition for tabletop manipulation*. Our approach is practical: we consider a typical interaction setup (Fig. 1) where the robot needs to accurately and robustly identify and localise objects in order to manipulate them, communicate about them or reason on their geometric properties and relations. Critically, the object recognition technique needs to be suitable for actual experimental work, including field experiments: it must be reasonably easy to deploy the system in a range of dynamic human environments, without having to rely on expensive or cumbersome physical sensors, or expensive computation. We also take a short to medium horizon: not all techniques we evaluate are commonly available yet, but all have the potential to be robust implementations in the near future.

This paper tries to remedy a lack of information on deployment details in HRI contexts: many traditional assessments do not report on practical considerations. We need to take into account many different factors. For example, how robust is the detection and pose recognition when there are frequent changes to the environment, such as varying backgrounds or changing lighting conditions.

Figure 1. A close proximity interaction setup, typically found in human-robot interaction and cognitive robotics scenarios. Key scene characteristics are usually constant: relatively small objects (e.g. largest side being less than 10 cm), presence of occlusions, limited working space, and the presence of both textured and texture-less objects.

In this paper we compare across three families of techniques. The first is techniques that rely on 2D images, from which we track a selection of points. Back projection on these points allow the estimation of an object's 6D position. The second family of methods use 3D templates. 3D objects are compared against a known point cloud to find the position and orientation of an object. The final family relies on techniques that do not use traditional vision techniques, for example RFID technology.

Surveys on Object Detection

As a cornerstone of many robotic applications, research on object recognition and localisation has been reviewed in numerous past literature surveys. These surveys typically focus on one family of techniques or algorithms, typically using synthetic datasets to quantitatively compare the performances of the state of the art. We summarise hereafter the main findings for each of the localisation techniques.

HAI '17, October 17–20, 2017, Bielefeld, Germany

ISBN 978-1-4503-5113-3/17/10…$15.00

DOI: https://doi.org/10.1145/3125739.3132593

Techniques based on 2D correspondences

When perceptual data consists of camera images, pre-stored templates of objects are often matched against the incoming video stream using 2D correspondence techniques. Li et al. [9] conducted a survey of visual feature detection. In the review they categorise these techniques based on the fundamental principle by which they detect features, such as edge, blob or corner detection. Feature detection methods vary in performance based on the application context, but among them feature based techniques such as A-KAZE, ORB and SURF are popular in object recognition and tracking contexts [5].

Techniques based on 3D correspondences

The increased availability and popularity of 3D cameras has driven the need for 3D object matching techniques. Diez et al. [6] performed a qualitative review of 3D registration techniques, in which a mapping is made between 3D images or a 3D templates and an image. They specifically reviewed a variety of detectors and descriptors for 3D registration. Descriptors and detectors attempt to minimise the number of points required before using such brute force techniques to perform accurate identification. Note that while these are used to select salient points, they nearly always end up using iterative closest point (ICP) algorithms, which find corresponding points between a template and an unknown object. The more points that are used, the more accurate the detection is, but using more points has an exponential impact on computational requirements.

Non vision-based techniques

Many other reviews also focus on technologies not relying on visual perception. RFID can be used for coarse localisation, and has been shown to have an accuracy of a few centimetres [13]. The techniques used in their review are meant for localisation within a room, while our focus is on techniques that work on the scale of under a metre, for example localising objects on a tabletop. But reduced distance holds potential for increased accuracy, as objects are nearer to the RFID readers. Mautz [10] conducted a wide survey of a number of indoor positioning techniques for a range of applications. Most of the techniques reviewed are localisation for navigation, and are not practical for use in a tabletop situation. However, among the suitable methods identified for the accuracy we require for tabletop recognition was magnetic technology, which is able to reach millimetre levels of precision. Hostettler et al. [8] look at using Anoto positioning technology to localise a robot. They concluded that using a printed pattern that they are able to position a robot with high accuracy and with robustness to lighting and occlusion conditions, the technology was only restricted by the size and quality of the sheets that could be printed with the pattern.

Approach and Methodology

We compare a number of existing implementations of a wide range of techniques for object and pose detection. We chose a selection of implementations based on availability, ability to process in real-time and that could be considered representative of that technology. Each of these methods was compared against the following criteria:

1. **Degrees of Freedom**: The degrees of freedom that the method is able to measure (position and/or orientation).
2. **Detection Stability**: How stable was the method of detection. Would an object be lost even if nothing was happening, or were false positives generated.
3. **Rotation Invariance**: Is the method able to track the object when it is rotated.
4. **Distance Invariance**: How much does the distance of the object affect the tracking for that method.
5. **Environment Interference**: Is the method able to cope with changes to the background and lighting.
6. **Occlusion**: Can the method detect objects that are being partially occluded by other objects from the perspective of the robot.
7. **Practical Use**: Any additional notes such as extra equipment required that may affect the usability of the system in an experiment.

Each method is briefly described. A table of results provides a side by side comparison of each implementation. Finally we discuss and provide recommendations on each method.

ASSESSMENT OF OBJECT DETECTION METHODS

Here we briefly describe each method we evaluated and their main weaknesses. Table 1 provides a summary of our results.

3D pose estimation from 2D images

These techniques use a standard 2D cameras. From this, image features are extracted that can be used to identify the object. These features can then be used to provide a 3D position by back projecting the 2D points to 3D reference points, using algorithms like 'perspective-n-point'(PnP) [7].

Fiducial markers

Fiducial markers look similar to 2D barcodes that can be printed out or displayed on a screen for detection. Several libraries provide 6D tracking of such markers, like the *chilitags* library [4].

Figure 2. Object with a fiducial marker, which allows it to be identified and tracked.

The tags are highly susceptible to occlusion, a small amount is enough to lose tracking. The markers require a flat surface to work, on irregularly shaped objects we get around this by attaching cubes (fig. 2).

Feature tracking

Three feature tracking methods were tested using the implementations provided by OpenCV[1]; SURF [2], A-KAZE [1] and ORB [12]. In each case an image is used as a target for the feature detection. These methods are classed as blob detection,

[1] http://opencv.org/

Method	Degrees of Freedom	Sta.	RInv.	DInv.	Env.	Occ.	Practical Use
2D w/ PnP							
Fiducial Markers	6D	Very High	Very High	High	Very High	Very Low	Markers on flat surfaces
A-KAZE	6D	Moderate	Very High	Low	Low	Moderate	
ORB	6D	Moderate	Very High	Low	Low	Moderate	
SURF	6D	Moderate	Moderate	Moderate	Low	Moderate	
Template Matching	6D	Very High	High	High	Low	Moderate	
Deep Learning (Faster R-CNN)	Planar	High	Very High	Very High	Very High	High	High Training Requirement
Depth Mapping							
ORK	6D	Very Low	High	High	High	Moderate	RGB-D Camera
Realsense SDK	6D	High	High	High	High	Moderate	RGB-D Camera
Non-Vision Based							
GaussSense	Planar w/ Rotation	Low	Very High	Very High	Very High	Very High	Sensor Board
ePawn	Planar w/ Rotation	Very High	Very High	Very High	Very High	Very High	Sensor Board

Table 1. Table showing a summary of the different object detection methods and their performance based on the criteria defined in section 1.3. Sta.: Detection Stability. RInv.: Rotation Invariance. DInv.: Distance Invariance. Env. Environment Interference. Occ.: Occlusion

which look for areas of pixels that are similar to each other but contrast their surroundings.

All three of these methods struggle with changing backgrounds, and did not handle varying distances well. Besides, computing the backprojection to obtain a 6D pose is generally difficult: as feature trackers select by themselves which features they choose to match, they are not known in advance. This makes it difficult to apply a PnP transformation to recompute 6D coordinates.

Template matching

Template matching, while a relatively old technique, was also considered; we tested using the implementation from OpenCV. An image is used as the target for template matching. This target image is then compared pixel by pixel against an image, and the strongest match is returned as a bounding box.

Multiple target images will be required per object to provide proper 6D pose estimation. Its greatest weakness is to varying backgrounds.

Deep Learning

Deep learning relies on the training of a neural network on a dataset of pictures. Here we used Faster R-CNN [11] to test Deep Learning. We used a pre-trained network[2] that was trained on the PASCAL VOC 2007 image dataset.

The network was unable to detect iconic versions of objects it had been trained on (fig. 3), so training would be required on the specific objects to be used as part of the experimental setup.

This method only provides bounding boxes of the objects, but these cannot be compared against a known object (an object could be small but near the camera or large but far away and we would be unable to determine the exact dimensions). This makes it difficult to provide a 6D estimation.

3D pose estimation from 3D sensor data

In recent years RGB-D cameras, which return 3D scene data in addition to a 2D image, have been widely used in HRI. The Microsoft Kinect technology or the Intel Realsense technology have proven particularly popular. Here we evaluate their software in the context of object localisation and pose reading. The techniques that we look do not require more than a tablet or laptop to process the data.

[2]https://github.com/smallcorgi/Faster-RCNN_TF

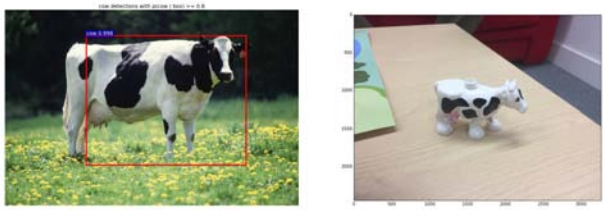


Figure 3. Images showing two pictures of cows, on the left a real cow that is detected by Faster R-CNN trained on the PASCAL VOC 2007 dataset, on the right an iconic toy cow that is missed.

Planar segmentation and iterative fitting

We evaluated "Tabletop" from the Object Recognition Kitchen (ORK)[3] implemented using ROS. Tabletop uses planar segmentation to separate the surface of a table and segment objects that are on top. These objects are then compared to a database containing meshes of known objects using simple iterative fitting (related to ICP[3]). This method performed well with different object rotations and scales, and was unaffected by a change in background. However this method generated too many false positives to be considered a stable option for close proximity human-robot interaction scenarios.

Intel Realsense tracking

In the Intel Realsense SDK[4], Object Tracking (C++) for the SR300 was used. This method relies on having a 3D mesh of the object, which it then used for matching. During our investigation we were unable to determine the exact method used by the Intel SDK as it has not been published (see discussion section). Objects were sometimes lost for no apparent reason and would need to be moved for them to be recognised again. This technique is able to handle a small amount of occlusion.

[3]http://wg-perception.github.io/object_recognition_core/index.html

[4]http://www.intel.co.uk/content/www/uk/en/architecture-and-technology/realsense-overview.html

Non-Vision Based Techniques

This section details methods that do not rely on the use of cameras, but instead the use of additional equipment.

Magnetic Field sensors

Magnetic Field sensors use one or more Hall effect sensors to read the position and orientation of a magnetic tag. We evaluated the GaussSense[5] solution, a small and affordable magnet sensor with a high degree of sensitivity. It is able to measure orientation and measures up to 3-4cm away from the sensor. It does however only cover a very small area. Many sensors would be required to cover a larger, the price may then become a consideration, with a 16x16cm board costing $350. GaussSense also requires the use of an Arduino to process the data received. However to distinguish between different tags requires an NFC tag.

NFC solutions

Several NFC sensors a can be combined into an NFC array, allowing for detection over a larger area. We evaluated the ePawn[6] mat, an NFC sensor board covering a 32x32cm area. The ePawn mat, using a 2D matrix of sensors, can locate a tag with millimetre accuracy. Using two tags in an object allows the calculation of orientation in the plane of an object. Tags themselves are 2cm in diameter so would be able to fit on or inside small objects. Tags only really work well while in contact with the mat. The prototype we evaluated currently costs €1400.

DISCUSSION AND RECOMMENDATIONS

Of all the 2D vision based techniques fiducial markers were probably the most reliable. However its sensitivity to occlusion means it is unsuitable for a study where the objects are frequently moved around by hand and placed behind other objects. Another challenge is often the attachment of fiducial markers onto objects: curved or irregular objects often prove challenging to attach the markers to. However, fiducial markers might bring benefits not offered by other technologies: the ease of displaying fiducial markers on a screen, or printing out markers, and the high accuracy it can provide, means that it is suitable for calibrating multiple cameras quickly in an experimental setup.

The feature tracking methods (A-KAZE, ORB and SURF) all have issues with dynamic backgrounds, which is an issue when the camera is not static or when subjects in the interaction are in view. It should be noted that the objects being used for this assessment were all relatively simple toys, which lacked rich texture. These methods may perform better on other, more textured, objects, but it may still require combining these methods with other algorithms to get a truly robust detection system.

Template matching, while relatively old, was among the most robust of the 2D methods. To provide a 6D pose estimation however this method will require a lot of templates to compare against. Therefore this method will not scale well with multiple objects. It may be better to use this method to increase the stability of other techniques where it could be used for foreground selection.

The Faster-RCNN that we tested can only provide a bounding box for our objects, this means we cannot get a full 6D pose estimation with this technique alone. However its reliability means that it could be very useful as a foreground selection technique to be used in a pipeline with other methods. Recent research looks into using a CNN that is able to handle 3D pose estimation [14], but it is unlikely that a training set for specific experimental requirements exist as these networks are only just emerging. The process of generating the required training data and then training the network is a process that potentially requires months of work before being usable in an experiment.

The implementation of tabletop in ORK provided too many false positives to be feasible for use in our future studies. However we only tried one camera, the Intel SR300. Other hardware or updates to software drivers may increase performance. By making use of the planar segmentation part of the process it would be possible to subtract the background for use in other detection methods, causing this to no longer be an issue for those methods which struggle with varying backgrounds.

The Intel Realsense SDK performed better with a lot higher stability compared to ORK. However the issue where it would sometimes lose an object while not common is still enough to cause issues in a study. This however is probably the best method available if it is a requirement to track objects while they are being moved. We were unable to find the exact technique that Intel Realsense used, as it has not been published, but due to its performance it was still included in this review. It appears to identify contours in the object before we assume using ICP to match these points to the points of objects stored in the database.

None of the vision based techniques were fully capable of performing the required level of object recognition in a practical tabletop setting. However a pipeline of techniques has the potential to overcome the weaknesses that are shown with just a single method. For instance the 2D techniques could be used to provide a bounding box and classification of the object, allowing a 3D technique to provide precision depth and pose information.

The GaussSense magnetic sensor performs well when tracking a single object. However an NFC module is required to be able to distinguish between multiple objects. For this reason it would be recommended to just use an NFC sensor when using multiple objects.

The ePawn NFC mat is probably the best method reviewed here for use in object recognition with tabletop manipulation. Its downside is that it cannot provide full 6D pose estimation, and the need for additional sensor equipment in the form of a RFID matrix. It is however suitable for many cases where objects need to be tracked, and potential interactions can be shaped around this limitation. NFC also has an advantage of being a known and reliable technique, as it used widely in contactless technology, such as debit cards and key fobs.

[5]http://gausstoys.com/
[6]http://epawn.fr/

ACKNOWLEDGEMENTS

This work was supported by the EU Horizon 2020 L2TOR project (grant 688014) the and EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227).

REFERENCES

1. Pablo F Alcantarilla and T Solutions. 2011. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell* 34, 7 (2011), 1281–1298.
2. Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. *Computer vision–ECCV 2006* (2006), 404–417.
3. Paul J Besl and Neil D McKay. 1992. Method for registration of 3-D shapes. In *Robotics-DL tentative*. International Society for Optics and Photonics, 586–606.
4. Quentin Bonnard, Séverin Lemaignan, Guillaume Zufferey, Andrea Mazzei, Sébastien Cuendet, Nan Li, Ayberk Özgür, and Pierre Dillenbourg. 2013. Chilitags 2: Robust Fiducial Markers for Augmented Reality and Robotics. (2013). `http://chili.epfl.ch/software`
5. Hsiang-Jen Chien, Chen-Chi Chuang, Chia-Yen Chen, and Reinhard Klette. 2016. When to use what feature? SIFT, SURF, ORB, or A-KAZE features for monocular visual odometry. In *Image and Vision Computing New Zealand (IVCNZ), 2016 International Conference on*. IEEE, 1–6.
6. Yago Diez, Ferran Roure, Xavier Lladó, and Joaquim Salvi. 2015. A qualitative review on 3D coarse registration methods. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 45.
7. Martin A Fischler and Robert C Bolles. 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24, 6 (1981), 381–395.
8. Lukas Hostettler, Ayberk Özgür, Séverin Lemaignan, Pierre Dillenbourg, and Francesco Mondada. 2016. Real-time high-accuracy 2D localization with structured patterns. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 4536–4543.
9. Yali Li, Shengjin Wang, Qi Tian, and Xiaoqing Ding. 2015. A survey of recent advances in visual feature detection. *Neurocomputing* 149 (2015), 736–751.
10. Rainer Mautz. 2012. Indoor positioning technologies. (2012).
11. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.
12. Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2564–2571.
13. T Sanpechuda and L Kovavisaruch. 2008. A review of RFID localization: Applications and techniques. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on*, Vol. 2. IEEE, 769–772.
14. Paul Wohlhart and Vincent Lepetit. 2015. Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3109–3118.

How to Manage Affective State in Child-Robot Tutoring Interactions?

Thorsten Schodde
CITEC, Bielefeld University
Bielefeld, Germany
tschodde@techfak.uni-bielefeld.de

Laura Hoffmann
CITEC, Bielefeld University
Bielefeld, Germany
lahoffmann@techfak.uni-bielefeld.de

Stefan Kopp
CITEC, Bielefeld University
Bielefeld, Germany
skopp@techfak.uni-bielefeld.de

***Abstract*—Social robots represent a fruitful enhancement of intelligent tutoring systems that can be used for one-to-one tutoring. The role of affective states during learning has so far only scarcely been considered in such systems, because it is unclear which cues should be tracked, how they should be interpreted, and how the system should react to them. Therefore, we conducted expert interviews with preschool teachers, and based on these results suggest a conceptual model for tracing and managing the affective state of preschool children during robot-child tutoring.**

I. Introduction

The use of robots for educational purposes has increasingly moved into focus in recent years. One rationale is to enable individually adapted one-to-one teaching for weaker students, which can hardly be provided in regular classrooms. This idea already underlay educational on-screen applications like intelligent tutoring systems (ITSs). Physically present social robots are expected to bring an additional quality to the learning interactions, similar to co-present teacher-child or child-child interaction, which can make the tutoring experience more effective. Indeed, a recent study showed that students' learning performance increases up to 50% if a social robot was included compared to a classical on-screen media learning [1].

One of the main challenges for robot tutors is to identify the learner's internal states, e.g., whether she is following, distracted, or losing motivation. Yet, recognizing and reacting to these cognitive and affective states is vital to keep the learner engaged and to foster learning. In previous work, we developed an approach to dynamically adapt robot tutoring to the changing pedagogical state of the learner [2]. There, the skill mastery of the student is kept track of inferentially using Bayesian Knowledge Tracing, which enables the robotic tutor to choose the to-be-addressed skill and difficulty of the next task accordingly. This way the model works to keep the child in the "zone of proximal development" [3], which can lead to a feeling of flow, motivation and better learning [4], [5].

However, this approach lacks "emotional intelligence" [6]. Successful human teachers not only teach the curriculum according to the learner's knowledge state, but also manage the affective states of children. Studies have shown that affective states like curiosity, interest, flow, joy, boredom, frustration and surprise can influence learner's problem-solving abilities, and affect task engagement and learning motivation [7]. Further, such states are found to influence cognitive processes like long-term memorizing, attention, understanding, remembering, reasoning, decision-making and the application of knowledge in task solving [4], [8]. It is thus not surprising that good human tutors are sensitive to learners' vocal (e.g., intonation) or nonvocal behavior (e.g., facial expression, body language) [9]. Technical systems are also increasingly able to recognize most of these cues - albeit sometimes in a quite rudimentary way. However, little attention has been paid to the question how a robot should interpret and respond to the affective state of a learner during tutoring with the needed flexibility and adaptiveness [10], [11].

In this paper we present steps towards a model for tracing and managing the affective state of preschool children in second language tutoring interactions with a robot tutor. This model is based on pedagogical knowledge about children's affective states during actual robot-child tutoring gathered through expert interviews with preschool teachers. This knowledge comprises information about which affective states are relevant, from which features they can be tracked and, finally, how to react to them appropriately as a tutor. It lends itself to a decision-theoretic affective state tracing model that can be combined with our previously developed adaptive knowledge tracing approach. The following section discusses previous work on affect detection and affective tutoring systems. Afterwards we present the procedure and results of the conducted expert interviews. Finally, we discuss how these findings can be incorporated into a conceptual model that enables the recognition of and reaction to changes in children's affective states.

II. Related Work

A. Affect Detection

A lot of work has been done on affect recognition based on different modalities. One widely used approach is the analysis of facial expressions to detect the affective state of a user [12]. Often, classifiers are trained on "very expressive and played" emotions, making their applicability to real-world interactions questionable. In fact, the accuracy of emotion detection based on facial features is often low in real-world applications. Furthermore, the recognition rate is strongly dependent on the expressiveness of each target.

An alternative approach is the detection of affect from the user's voice [13]. Classifiers based on voice analysis are

trained on datasets of spontaneous speech, so that they are more suitable for real-world applications. With regard to robot-child tutoring, affect detection through speech analysis is, however, difficult because speech input is often not included as speech recognition for children has a low accuracy [14]. Other attempts have been made to detect the affective state through analyzing written text [15]. This approach includes, for instance, analyzing the usage of adjectives and adverbs. But in most natural interactions humans do not write text, and preschool children are usually not able to read and write.

A broader approach for affective state detection is the tracking of the whole body posture and movements by using a body pressure mat laying on a seat [16], or using a Microsoft Kinect [17]. A limitation is that the use of a body pressure mat assumes that the user remains on a seat and cannot move around. The Kinect, however, allows the user to move around, but may have problems in detecting smaller events like small postural shifts. Also, approaches based on human physiology have been adopted. In this realm, measures such as ECG, EEG, EMG [18], [19], and brain imaging [20] have been applied to "read" the affective state from the user's body. The results of these methods are promising, however the applicability of such obtrusive approaches (e.g., wires and patches on the body) in tutoring interactions with children is clearly limited.

In sum, all of these approaches have their field of use, but also their limitations. In contrast, multi-modal approaches have been studied to overcome these limitations and to increase accuracy of the detection. A lot of combinations exist, e.g., facial expressions and voice [21], facial expression, voice and body posture [22], facial expressions, body postures and context dependent activity logs [23], or speech and text [24]. Such systems demonstrated that a multi-modal approach to detect affective states results in higher accuracy rates.

B. Affective Tutoring Systems (ATSs)

Since the technical progress yields new possibilities to make use of the affective state in tutoring interactions, a lot of systems have been extended with such a module. Shen et al. [25], for instance, used physiological signals for affect detection and then guided the learning interaction by different affective strategies. Their results demonstrated the superiority of an emotion-aware over a non-emotion-aware system with a performance increase of 91% .

Alexander et al. [26] developed an affect-detecting ITS including a virtual agent for primary school students. The affective state is detected by analyzing the facial expressions of the student and serves as the basis for a case-based selection of the next tutoring actions. The case-based rules have been informed by an observational study of human tutors. In a study conducted in a primary school, where children had to solve mathematical equations, the use of their affective system showed a significant increase of the students' performance as compared to a control group without affective support.

The "Affective AutoTutor" system [27] can automatically detect boredom, confusion, frustration and neutral affect by monitoring conversational cues and discourse features along with gross body language and facial features. Cues provided by each channel are combined to select a single affective state, based on which AutoTutor responds with empathic, motivational, or encouraging dialog-moves and emotional displays. Evaluations showed that this systems is able to support learners not only in acquiring knowledge, but also in using it in transfer tasks later on. Recently, Goren et al. [28] incorporated affect detection via facial expressions in robot-child tutoring. In a study with preschool children they showed that their system personalized its policy over the course of training, and that children who interacted with the personalized robot showed increased long-term positive valence as compared to a control group without personalization.

Taken together, the findings from earlier approaches suggest the inclusion of affect detection in robot-child tutoring. Most affect detectors are trained on specifically annotated data to identify the important cues for each affective state. For example, the emotion classifier "Affectiva Affdex" [29] is trained on more than 5 million human faces to classify facial expressions. Strategies for how to respond to those states are usually based on observational studies of the reactions of a human tutor to the behavior of a student [30]. We adopt this approach here, too, with the aim of building a model that enables a robot to detect changes in children's learning-relevant affective states and to react to these changes appropriately. For this, child-robot interaction specific knowledge is necessary that could be best gathered from experts in reading and managing the affective states of young children in tutoring interactions, namely, preschool teachers.

III. Empirical Basis

With the aim of answering the questions, which affective states occur and are important during robot-child tutoring, and how they can be detected based on the observation of a child, a qualitative approach was chosen. We used video recordings from a previous study in kindergarten and interviewed five preschool teachers on their perception and interpretation of the children's behavior.

A. Participants

A total of five female preschool teachers were invited and interviewed as experts. They were between 36 - 61 years old ($M = 48.6; SD = 8.16$) and had a working experience from 16 to 42 years ($M = 29; SD = 8.88$).

B. Materials

With the objective of allowing the experts to observe children during robot-child tutoring in a controlled manner, video recordings from an interaction study were used. They were presented and discussed during face-to-face interviews with one interviewer. In total, video recordings of eight different children (4 female, 4 male), which varied in their level of activity and expressiveness when facing the robot, were chosen. The decision was taken to ensure that individual difference are considered in spite of the small samples. The recordings were taken in the realm of a separate study in Dutch

Fig. 1. Screenshot from one of the videos shown to the experts during the interview. The learning interaction is displayed from two perspectives.

preschools were children were tutored to learn animal names in a foreign language by means of a "I spy with my little eye..." game with a Nao robot. Here, up to four images of animals were displayed on a tablet screen, while the robot is referring to one of them using a Dutch description and the English name of the animal [31]. To choose the animal the robot mentioned, the children had to tap on the picture on the tablet. Two camera perspectives were recorded and presented to the experts to allow a frontal view on the child, but also a landscape view from the side on the whole experimental setup which includes the robot, the tablet and the child (see Fig. 1).

C. Procedure

At the beginning of each interview session, the participants were informed about the purpose and the procedure of the interview and signed an informed consent that their voice was recorded. They were instructed that they should judge the behavior and related affective state of children, which are presented in video recordings. First, a small example video was presented, which had to be commented by the experts to make sure the task was clear. Then, the interviewer started the video on a laptop and asked the expert to comment on the child's behavior and state. After each video (one video relates to one child) the interviewer asked how the experts would react to negative changes in the child's state, e.g., if they recognize a lack of attention, and how this could be realized with a robot. At each point in time, the interviewees were allowed to pause the video and go back to review a scene. Each expert discussed a total of four videos with the interviewer. Afterwards they were thanked for their participation and dismissed.

D. Analyses and Results

The whole interview session were recorded by means of a computer microphone, and a screen capture tool to synchronize the comments with the video recording that was played at the time. The recordings were afterwards transcribed to enable detailed content analyses of the experts' comments. The transcripts were then analyzed regarding the following research questions:

TABLE I
CHILDREN'S STATES AND RELATED CUES

Meta-level State	**State Interpretation**	**Behavioral Cue**	**n***
Engagement	Concentration/ Thinking	eye contact	5 (4)
		sit still	2 (2)
		hand to head	4 (3)
	Involvement/ Activity	mimic robots gestures	2 (2)
		answer verbally	1 (1)
		nodding	1 (1)
		head-shaking	1 (1)
	Expressive/Proud	smiling	7 (4)
		thumb up	1 (1)
		raise fist	1 (1)
Disengagement	Inattentiveness/ Distraction	rub eyes	2 (1)
		grimace	4 (4)
		gaze away	7 (4)
		turn away (whole body)	10(4)
		move position (stand up, lay down)	2 (2)
	Boredom/ Impatience	support the head with hand(s)	3 (2)
		move the head from left to right	2 (2)
		undirected finger tapping	4 (3)
		gaze away	2 (1)
		move position (stand up, lay down)	6 (4)
Negative Engagement	Skepticism	tilt head	3 (3)
	Disinterest	frown	1 (1)
	Averseness	lower mouth corners	1 (1)

*n is the frequency of reference to a cue; the amount of children for which the cue was observed is noted in parentheses.

- RQ1: How do experts interpret the cognitive and emotional state of children during the robot-child tutoring lessons?
- RQ2: To which behavioral cues do they refer when they remark changes (e.g., in the childs level of attention)?
- RQ3: How would the experts react to changes in the children's engagement from the perspective of the robot?

According to the experts descriptions of the children's states, categories of states were derived. As listed in Table I, the childrens states can be classified into states of engagement, disengagement, and negative engagement, on a meta level (RQ1). Engagement is composed of concentration and thinking, activity and involvement, as well as expressiveness. If a child kept eye contact with the robot and tablet, and sit still, the experts interpreted their behavior as concentrated and engaged. If they mimicked the gestures the robot made, or answered verbally or nonverbally (e.g., nodding, head-shaking), they were also described as involved and thus engaged in the interaction. Likewise, expressive behaviors as smiling, or showing a thumb up were interpreted as a sign of engagement by the experts. On the other hand, behaviors that were interpreted

as signs of inattentiveness and distraction, or boredom, were regarded as indicators of disengagement. For instance, rubbing eyes, gazing away, or frequent changes of the seating position were interpreted as inattentiveness. Additionally, supporting ones head with the hands, undirected tapping with the fingers, and gazing away, were (among others, cf. Table I) named as remarkable behaviors that demonstrate boredom and disengagement. Finally, the category negative engagement contains negative states like skepticism and averseness. These states were related to frowning, lowering mouth corners, and head-tilt (RQ2).

Each interaction with the robot varied according to individual differences of the children (e.g., age, self-confidence). Hence, we counted for each behavioral cue, how many times it was mentioned by different experts for different children. If two experts observed a cue for one child as relevant it was counted as two; but if one expert mentioned one cue for one child several times it was counted as one. To reflect on the occurrence of the cues over different children, it was further listed for how many different children the cue was observed (see Table I numbers in parentheses).

The results indicate that eye contact ($n = 4$ children), smiling ($n = 4$), and self-touches to the head ($n = 3$) were interpreted as a sign of engagement for multiple children in the video recordings. Regarding disengagement, making grimaces ($n = 4$), gazing away ($n = 7$), turning away ($n = 4$), moving the position ($n = 2$), and finger tapping ($n = 3$) were observed across several children. As a sign of negative engagement, head tilt was for several children ($n = 3$) interpreted as showing skepticism. Instead, giving verbal answers, nodding, head-shake, eye rub, frowning, and lowered mouth corners were only addressed for one child, respectively, and appear hence less informative. Note that the counts refer to the spontaneous mention of the cue per child and that the cues were overall mentioned repeatedly over the course of the interaction.

Furthermore, we asked the experts how they would intervene to keep children engaged in the interaction from the robots point of view (RQ3). Their suggestions were summarized into categories of potential actions to re-engage children in the tutoring with the robot (Table II).

Parts of the experts suggestions can be regarded as preventive strategies that can be employed in the interaction from the outset. These are general strategies to keep children engaged in an interaction as allowing multi-modal interactions (here: add speech) or more expressive robot behavior (e.g., gestures, movements). Beyond that, actions were mentioned that can be useful to re-engage children in an ongoing interaction after their engagement was lowered (repair actions, see Table II). The robot could for example suggest alternative activities to get the child's attention back (e.g., play a game). In some cases, it will even be necessary to stop the tutoring for a break according to the expert's opinions. Moreover, it was suggested that the difficulty of the task should be increased if signs of disengagement are recognizable.

TABLE II
POSSIBLE ACTIONS MENTIONED BY THE EXPERTS

Preventive actions	**Paraphrases**	**n***
Include verbal input	It would be more motivating for the child if it should talk to the robot (expert 2, video 2)	3
Heighten robot's activity (e.g., move head)	The interaction would be more engaging if the robot moves. (expert 2, video 2)	3
Repair actions		
React to the child's behavior/ give feedback	The robot should react to the behavior of the child, e.g., tell him/her to sit down again. (expert 5, video 1)	4
Change task difficulty	The task should increase in difficulty to get the childs attention back. (expert 1, video 3)	1
Include alternative activities (e.g., play a game; stand up)	The robot could ask the child to stand up and move around, so that he/she is ready to listen again afterwards. (expert 3, video 2)	4
Allow a break	A break or a continuation at another day could be helpful to get the attention back (expert 2, video 1)	2

*n is the amount of experts out of the 5 experts that mentioned the strategy.

E. Discussion

In summary, the analyses of the expert interviews revealed that preschool teachers agree on the interpretation of several child behaviors as signs of (dis-)engagement. The behavioral cues that were identified during robot-child tutoring were changes in gaze direction (eye contact versus gaze away), body posture (turn away, stand up, lay down), or facial expressions (smiling). These cues that have been identified can be used to narrow down the feature space in affective state recognition. We note, though, that the small amount of video samples restricts the significance of our findings. However, a frequent, independent naming of the most relevant cues by different experts for different children points to the importance of these cues for detecting the affective state of children. Interestingly, the majority of these cues can be recorded by means of non-obtrusive technologies (e.g., video cameras, Microsoft Kinect) and can be extracted using existing tools (e.g., Affdex, see above). Building on this, the following section lays out a conceptual approach to interpret and respond to changes in the child's state during robot-child tutoring interactions.

IV. AFFECTIVE STATE MANAGEMENT MODEL

A. Tracing the Affective State

The first step is to combine the different cues mentioned in Section III into higher-level states and to trace them over time. As a first approach, this can be achieved using a naive Bayesian classifier that determines the hidden internal state E that is assumed to independently cause cues $C_1, C_2, ..., C_n$. Since cues need to be integrated into coherent belief updates over time, the corresponding belief must be updated every time step according to a dynamic Bayesian model $P(E^{t+1}|C_i^{t+1}, E_t)$.

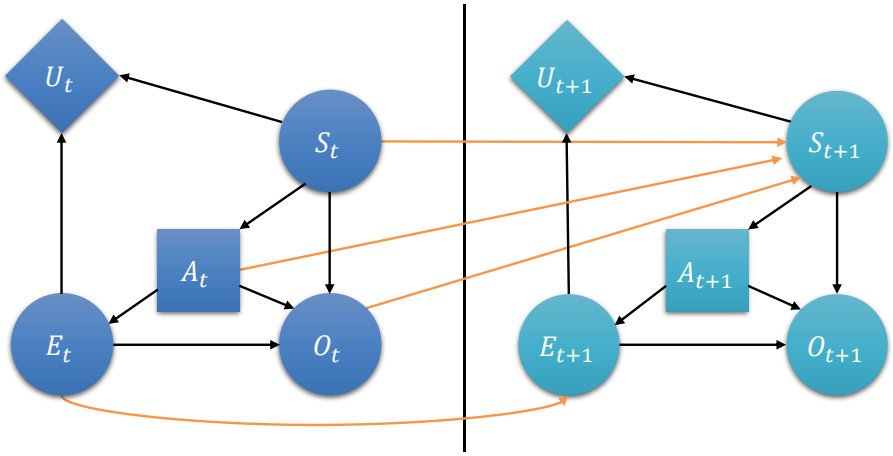


Fig. 2. Here the adaptive Bayesian Knowledge Tracing model is shown, consisting of the belief regarding the mastery of a skill S_t, the observation (response) O_t to an action A_t, the affective state E_t of the learner and the expected value U_t of a chosen chain of actions.

Variables E and C_i are directly based on the results of the expert interviews. We focus on the most reliable and explicit cues that can be tracked with current technology. Thus we base the model on those cues that were frequently mentioned for several children (cf. Table I). Since most cues from the negative engagement group were only mentioned once, and "head tilt" is difficult to track due to the danger of mixing it up with moving the head from side to side (from the disengagement group), we focus on signs of engagement and disengagement in the first stage of the model's development. Engagement and disengagement can be regarded as opposing poles on a continuum of engagement. Hence, we combine them into the meta state variable E_t that is called *interaction engagement*. Cues that were identified as indicating engagement will have a positive effect on this state, while all cues related to disengagement will have a negative impact.

B. Managing the Affective State

After computing the belief update for interaction engagement, the next step is to determine whether and how the robot tutor should act. To this end, we include the belief variable E into our previously developed approach based on Bayesian Knowledge Tracing [2] (see Fig. 2). According to this model, the belief over the learners mastery of a certain skill S_t explains the observed answer O_t to a given teaching-task A_t selected to address this skill S_t. We add the state variable E_t as well as an utility value U_t, which represents the expected value of a chosen chain of pedagogical and affective actions. E_t is assumed to influence the students answer to a task, e.g. if the student is disengaged there may be a higher probability of observing a wrong answer as she may not have understood the task description. This information will also affect the belief update for the currently addressed skill, so that a wrong answer will have a lower impact when the student is disengaged.

Although experts' agreed on the identification of the behavioral cues, the interpretation of these cues should be regarded carefully since one behavior could have distinct meanings depending on the situation and the specific child. For the realization of a general model, the expert information is useful to determine which cues are relevant to look at as a starting point. A final system must, however, be able to adapt to specific variations in the child and the situation.

Next, we need to extend the action space of A_t to actions that manage the affective state, in addition to the already present actions of addressing a certain skill with a particular task. This allows evaluating and weighing both options, teaching a skill or managing the affective state of a student. Still, the main goal is to find an action (or action sequence) from which the child will learn the most. Since the model is a Dynamic Bayesian Decision Network, this evaluation can be carried out across several time steps, where each additional time step lowers the utility gained on the basis of the increase of the skill belief. Hence, the system can decide whether it is more beneficial to first raise interaction engagement, before teaching the next skill, or the other way around.

Again, we based our selection of actions to manage affective state on the results of the expert interviews (cf. Table II). We consider only the repair actions here, out of which the change of task difficulty is already implemented in the model. Three other actions remain, which could be useful to re-engage a child in the interaction: First, directly addressing the child's behavior, e.g., urge to sit down again or ask for attention; secondly, using alternative tasks or activities to provide a more variable interaction, e.g., ask to move around or to play a game; finally, if the interaction engagement drops significantly, the robot can propose a break and the interaction can be resumed later. All of these behaviors can be immediately included in the model as well as the robot's behavior repertoire. Note, however, that the conditional probabilities $P(E|A)$ ans $P(O|A, E, S)$ need to be defined heuristically as long as sufficient interaction data is not available.

V. Summary

The present paper addressed the importance of coping with a learner's affective state during preschool child-robot tutoring. While the automatic recognition of cues seems to

be within reach with today's technology, we are still lacking a model of which affective states are most relevant in such learning interactions, how they can be recognized, and how they should be responded to by the robot tutor. To tackle this problem, expert interviews with preschool teachers have been conducted to identify children's affective states that are relevant during robot-child tutoring. The results suggest that different categories of engagement states seem to be most important, and that experts recognize and address those states in interaction. The findings from the interviews are currently used to inform the implementation of a computational model for tracing and managing the affective and cognitive state of a child learner with a robot tutor. To this end, we have laid out how to extend a previously developed knowledge-tracing and decision-making model based on a dynamic Bayesian Decision Network. The combined model will allow for finding an action policy that combines informative and affective actions of a robot tutor to manage the internal states (both, cognitive and affective) of a child learner more thoroughly, and to ensure an optimal course of learning.

ACKNOWLEDGMENT

We thank our colleagues from Tilburg University who provided the videos for the interviews. This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG), and by the L2TOR (www.l2tor.eu) project, grant number: 688014, and by the BabyRobot (www.babyrobot.eu) project, grant number: 687831, both supported by the EU Horizon 2020 Program.

REFERENCES

[1] J. Kennedy, P. Baxter, and T. Belpaeme, "The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning," in *Proceedings of ACM/IEEE HRI 2017*. ACM, 2015, pp. 67–74.

[2] T. Schodde, K. Bergmann, and S. Kopp, "Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making," in *Proceedings of ACM/IEEE HRI 2017*. ACM Press, 2017, pp. 128–136.

[3] L. Vygotsky, *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press, 1978.

[4] S. Craig, A. Graesser, J. Sullins, and B. Gholson, "Affect and learning: an exploratory look into the role of affect in learning with autotutor," *Journal of educational media*, vol. 29, no. 3, pp. 241–250, 2004.

[5] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes, "Information-seeking, curiosity, and attention: computational and neural mechanisms," *Trends in cognitive sciences*, vol. 17, no. 11, pp. 585–593, 2013.

[6] N. Thompson and T. J. McGill, "Affective tutoring systems: enhancing e-learning with the emotional awareness of a human tutor," *International Journal of Information and Communication Technology Education*, vol. 8, no. 4, pp. 75–89, 2012.

[7] N. Schwarz, "Emotion, cognition, and decision making," *Cognition & Emotion*, vol. 14, no. 4, pp. 433–440, 2000.

[8] B. Lehman, S. DMello, and N. Person, "The intricate dance between cognition and emotion during expert tutoring," in *Intelligent Tutoring Systems*. Springer, 2010, pp. 1–10.

[9] S. Petrovica and M. Pudane, "Simulation of affective student-tutor interaction for affective tutoring systems: Design of knowledge structure," in *Proceedings of EET 2016*, vol. 7, 2016.

[10] S. DMello, N. Blanchard, R. Baker, J. Ocumpaugh, and K. Brawner, "I feel your pain: A selective review of affect-sensitive instructional strategies," *Design Recommendations for Intelligent Tutoring Systems*, vol. 2, pp. 35–48, 2014.

[11] R. A. Sottilare, J. A. DeFalco, and J. Connor, "A guide to instructional techniques, strategies and tactics to manage learner affect, engagement, and grit," *Design Recommendations for Intelligent Tutoring Systems*, vol. 2, pp. 7–33, 2014.

[12] B. McDaniel, S. D'Mello, B. King, P. Chipman, K. Tapp, and A. Graesser, "Facial features for affective state detection in learning environments," in *Proceedings of the CogSci 2007*, vol. 29, no. 29, 2007.

[13] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs." in *Proceedings of Interspeech 2006*, 2006.

[14] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child speech recognition in human-robot interaction: evaluations and recommendations," in *Proceedings of ACM/IEEE HRI 2017*. ACM, 2017, pp. 82–90.

[15] J. H. Kahn, R. M. Tobin, A. E. Massey, and J. A. Anderson, "Measuring emotional expression with the linguistic inquiry and word count," *The American journal of psychology*, pp. 263–286, 2007.

[16] S. D'Mello and A. Graesser, "Automatic detection of learner's affect from gross body language," *Applied Artificial Intelligence*, vol. 23, no. 2, pp. 123–150, 2009.

[17] D. McColl and G. Nejat, "Affect detection from body language during social hri," in *21st IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2012, pp. 1013–1018.

[18] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *Proceedings of ICME 2005*. IEEE, 2005, pp. 940–943.

[19] O. Villon and C. Lisetti, "A user-modeling approach to build user's psycho-physiological maps of emotions using bio-sensors," in *15th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2006, pp. 269–276.

[20] M. H. Immordino-Yang and A. Damasio, "We feel, therefore we learn: The relevance of affective and social neuroscience to education," *Mind, brain, and education*, vol. 1, no. 1, pp. 3–10, 2007.

[21] A. Esposito, "Affect in multimodal information," in *Affective Information Processing*. Springer, 2009, pp. 203–226.

[22] T. Bänziger, D. Grandjean, and K. R. Scherer, "Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (mert)." *Emotion*, vol. 9, no. 5, p. 691, 2009.

[23] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of MM 2005*. ACM, 2005, pp. 677–682.

[24] I. Arroyo, D. G. Cooper, W. Burleson, B. P. Woolf, K. Muldner, and R. Christopherson, "Emotion sensors go to school." in *Proceedings of AIED 2009*, vol. 200, 2009, pp. 17–24.

[25] L. Shen, M. Wang, and R. Shen, "Affective e-Learning: Using emotional data to improve learning in pervasive learning environment related work and the pervasive e-learning platform," *Educational Technology & Society*, vol. 12, pp. 176–189, 2009.

[26] S. Alexander, A. Sarrafzadeh, S. Hill *et al.*, "Easy with eve: A functional affective tutoring system," in *Workshop on Motivational and Affective Issues in ITS*. Citeseer, 2006, pp. 5–12.

[27] S. D'mello and A. Graesser, "Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back," *Interactive Intelligent Systems*, vol. 2, no. 4, p. 23, 2012.

[28] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Proceedings of 30th AAAI Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 3951–3957.

[29] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, "Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit," in *Proceedings of CHI 2016*. ACM, 2016, pp. 3723–3726.

[30] S. Alexander, A. Sarrafzadeh, and S. Hill, "Foundation of an affective tutoring system: Learning how human tutors adapt to student emotion," *International journal of intelligent systems technologies and applications*, vol. 4, no. 3-4, pp. 355–367, 2008.

[31] J. de Wit, T. Schodde, B. Willemsen, K. Bergmann, M. de Haas, S. Kopp, E. Krahmer, and P. Vogt, "Exploring the Effect of Gestures and Adaptive Tutoring on Childrens Comprehension of L2 Vocabularies," in *Proceedings of the Workshop R4L at ACM/IEEE HRI 2017*, 2017.

The Free-play Sandbox: a Methodology for the Evaluation of Social Robotics and a Dataset of Social Interactions

Séverin Lemaignan, Charlotte Edmunds, Emmanuel Senft, Tony Belpaeme
Plymouth University
Plymouth, United Kingdom
Email: firstname.lastname@plymouth.ac.uk

December 8, 2017

Abstract

Evaluating human-robot social interactions in a rigorous manner is notoriously difficult: studies are either conducted in labs with constrained protocols to allow for robust measurements and a degree of replicability, but at the cost of ecological validity; or *in the wild*, which leads to superior experimental realism, but often with limited replicability and at the expense of rigorous interaction metrics.

We introduce a novel interaction paradigm, designed to elicit rich and varied social interactions while having desirable scientific properties (replicability, clear metrics, possibility of either autonomous or Wizard-of-Oz robot behaviours). This paradigm focuses on child-robot interactions, and builds on a sandboxed free-play environment. We present the rationale and design of the interaction paradigm, its methodological and technical aspects (including the open-source implementation of the software platform), as well as two large open datasets acquired with this paradigm, and meant to act as experimental baselines for future research.

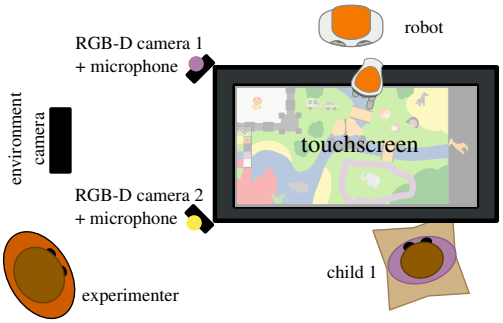


Figure 1: The free-play social interactions sandbox: two children interact in a free-play situation, by drawing and manipulating items on a touchscreen. Children are facing each other and sit on cushions. Each child wears a bright sports bib, either purple or yellow, to facilitate later identification.

1 The challenges in evaluating social interactions

1.1 Studying social interactions

Studying social interactions requires a social *situation* that effectively elicits interactions between the participants. Such a situation is typically scaffolded by a social task, and consequently, the na-

ture of this task influences in fundamental ways the kind of interactions that might be observed and analysed. In particular, the socio-cognitive tasks commonly found in the literature of experimental psychology (and HRI) often have a narrow focus: because they aim at studying one (or a few) specific social or cognitive skills in isolation and in a controlled manner, these tasks are typically simple and highly constrained (for instance, an object handover task; a perspective-taking task with cubes, etc.). While these focused endeavours are important and necessary, we – as a community – also acknowledge that these interaction scenarios do not reflect the complexity and dynamics of real-world interactions Baxter et al. (2016), and we certainly observe a strong trend within our community towards capturing, interpreting and acting upon the rich set of naturally-occurring social interactions.

Specifically, we believe that further progress in the study of human-robot interactions should be scaffolded by socio-cognitive challenges that:

- are long enough and varied enough to elicit a large range of interaction situations;
- foster rich multi-modal interaction, such as simultaneous speech, gesture, and gaze behaviours;
- are loosely directed, to maximise natural, non-contrived behaviours;
- evidence complex social dynamics, such as rhythmic coupling, joint attention, implicit turn-taking;
- include a certain level of non-determinism and unpredictability.

The challenge lies in designing a social task that exhibits these features *while maintaining 'good' scientific properties* (repeatability, replicability, robust metrics) as well as good practical properties (not requiring unique or otherwise very costly experimental environments, not requiring very specific hardware or robotic platform, easy deployment, short enough experimental sessions to allow for large groups of participants).

In this paper, we introduce such a task, designed to elicit rich, complex, varied social interactions while being well suited for interactions with robots and supporting rigorous scientific methodologies.

1.2 Social play

Our interaction paradigm is based on free and playful interactions (free play) in a *sandboxed* environment: while the interaction is free (participants are not directed to perform any particular task beyond playing), the activity is both *scaffolded* and *constrained* by the setup mediating the interaction (essentially, a large table-top touchscreen). Participant engage in open-ended and non-directive play situations, yet sufficiently well defined to be reproducible and practical to record and analyse.

This initial description frames the socio-cognitive interactions that might be observed and studied: playful, dyadic, face-to-face interactions. While gestures and manipulations (including joint manipulations) play an important role in this paradigm, the participants do not typically move much during the interaction. Because it builds on play, this paradigm is also naturally suited to the study of child-child and child-robot interactions.

The choice of a playful interaction is supported by the wealth of social situations and social behaviours that *play* elicits. Most of the research in this field builds on the early work of Parten who established five *stages of play* Parten (1932), corresponding to different stages of development, and accordingly associated with typical age ranges:

1. **Solitary (independent) play**, age 2-3: Playing separately from others, with no reference to what others are doing.
2. **Onlooker play**, age 2.5-3.5: Watching others play. May engage in conversation but not engage in doing. True focus on the children at play.
3. **Parallel play** (adjacent play, social coaction), age 2.5-3.5: Playing with similar objects, clearly beside others but not with them (near but not with others.)
4. **Associative play**, age 3-4: Playing with others without organization of play activity. Initiating or responding to interaction with peers.
5. **Cooperative play**, age 4+: Coordinating one's behavior with that of a peer. Everyone has a role, with the emergence of a sense of belonging to a group. Beginning of "team work."

These five stages of play have been extensively discussed and refined over the last century, yet remain remarkably widely accepted as such. It must be noted that the age ranges are only indicative. In particular, most of the early behaviours still occur at times by older children.

Interestingly, these five stages can been looked at from the perspective of HRI as well. They certainly evoke a roadmap for the development of human-robot social interactions.

2 The Free-play Sandbox

2.1 Task

We have designed a new experimental task, called the *free-play sandbox*, that is based on free play interactions. Pairs of children (4-8 years old) are invited to freely draw and interact with items displayed on an interactive table, without any explicit goal set by the experimenter (Fig. 1). The task is designed so that children can engage in open-ended and non-directive play, yet it is sufficiently constrained to be suitable for recording, and allows the reproduction of social behaviour by an artificial agent in comparable conditions.

The free-play sandbox follows the sandtray paradigm Baxter et al. (2012): a large touchscreen (60cm $\times$ 33cm, with multitouch support) is used as an interactive surface (*sandtray*). Two children play together by freely moving interactive items on the surface (Fig. 2). A background image depicts a generic empty environment, with different symbolic colours (water, grass, beach, bushes...). By drawing on top of the background picture, the children can change the environment to their liking. The players do not have any particular task to complete, they are simply invited to freely play. Importantly, they can play for as long as they wish (for practical reasons, we have limited the sessions to a maximum of 40 minutes in our own experiments, see Section 5).

Capturing all the interactions taking place during the play sessions is possible and practical with this setup. Even though the children will typically move a little, the task is fundamentally a face-to-face, spatially delimited, interaction, and as such simplifies the data collection. For instance, during our dataset acquisition campaign (120 children, more than 45h of footage), the children's faces were automatically detected in 98% of the recorded frames (see Section 5).

Figure 2: Example of a possible game situation. Items (animals, characters...) can be dragged over the whole play area, while the background picture can be painted over by picking a colour.

2.2 Applications

Child-Child Interaction The free-play sandbox provides the opportunity to observe children interacting in a natural way in an open but framed setup. As the system can run on a single computer platform it can easily be deployed in the 'wild', in places where the children naturally interact such as classroom. The quantity and thoughtfulness of information logged allows to keep a track of every interaction happening around the game.

These advantages combined with the openness of the task proposed make this setup a powerful tool to observe and quantify a large spectrum of social behaviours expressed by children when interacting in a natural environment (might be interesting to add a list here). The compactness of the system makes it easy to compare data from different locations.

Child-Robot Interaction This free-play sandbox provides the opportunity to explore child-robot interactions in this open, real world environment as shown in Figure 1.

Depending of the focus of the study, two modes of control for the robot are available. If the interest is on evaluating a specific robot behaviour, the robot can be autonomously controlled using inputs from the different sensors. This setup allows to explore the impact of different social behaviours

on the children independently of the 'game policy' controlling by the robot.

On the other hand, if the focus is on the child behaviour and the technical aspect is of a lower importance, the robot can be controlled by a human rather than an algorithm. This paradigm, where the robot is tele-operated to interact with a naive partner is called Wizard of Oz (WoZ) and is used in numerous studies to explore the psychologic side of HRI Riek (2012).

Deep Learning With the quantity of data logged and the high number of interaction achievable with the free-play sandbox, it supports the type of requirement for recent Machine Learning approaches such as deep learning. The similar position of the children in all interactions makes the combination of data from different interaction easier than other less compact systems.

From the information collected on the children, social behaviours can be extracted and used on a robot.

3 Implementation

The software-side of the free-play sandbox is entirely open-source[1]. It is implemented using two main frameworks: Qt QML[2] for the graphical interface of the game, and the *Robot Operating System* (ROS) for the modular implementation of the data processing and behaviour generation pipelines. The graphical interface interacts with the decisional pipeline over a bidirectional QML-ROS bridge that we have developed for that purpose.

Figure 3 presents the software architecture of the sandbox.

3.1 Interactive game

The interactive game (Fig. 3.1) is coded using QML, and displays a main background image on top of which items (animals, humans and objects) can be moved. The children can also use a drawing mode to create coloured strokes on a layer between the background and the items, which adds another layer of unconstrained interaction to the game (Figure 2). The game exposes the image of the background, the drawings, and the positions of the objects as ROS TF frames.

3.2 Sensing

Two Intel RealSense SR300 RGB-D cameras are mounted at fixed positions on the sandtray frame, with custom designed 3D-printed brackets that ensure that the cameras are oriented towards the children's face. Because the cameras are rigidly mounted onto the sandtray's frame, their accurate geometric transformations with respect to the sandtray screen are known. Combined with hardware calibration, it allows for accurate localisation of the children and in particular, children's faces. In addition to the images, both cameras can perform stereo audio recording. One ROS node per camera (Fig. 3.2) publishes on dedicated topics the audio and video streams.

A third 'external' (and non-calibrated) camera is usually used as well to record the environment of the experiment with a wider angle (*environment camera* in Figure 1).

3.3 Robot Control

As stated in section 2.2, a robot (Fig. 3.9) can act as play partner instead of one of the children. This robot can either be autonomous selecting actions based on the inputs provided by the sensors and the game or be controlled by a human in a Wizard of Oz fashion.

Autonomous The current implementation exposes a large number of information on the game and the state of the child that can be used in the robot controller. The position of every item is exposed as a TF frame, the background is segmented in zones of identical colors (Fig. 3.5), social element of the state the interaction are collected through the RGBD camera and the microphone facing the child. As visible on Figures 1 and 4, the camera covers the head of the child as well as most of the upperbody, and applying libraries such as DLib and OpenPose, the position of facial feature and skeleton of the child are extracted and can be used to obtain: head gaze, gaze and gestures such as pointing. All these inputs can be combined to provide

[1]Source code: `https://github.com/freeplay-sandbox/core`

[2]`http://doc.qt.io/qt-5/qmlapplications.html`

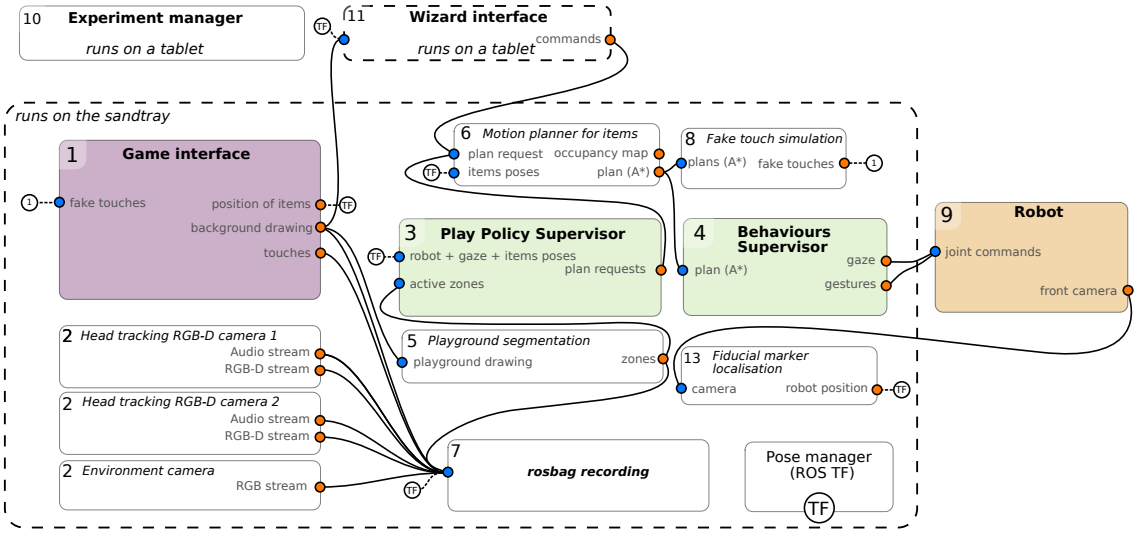


Figure 3: Software architecture of the free-play sandbox. Left (purple) nodes are connected to the sandtray (game interface (1) and camera drivers (2)). Nodes in the centre (green) implement the behaviour of the robot (play policy (3) and robot behaviours (4)). Several helper nodes are available, in particular, segmentation of the children drawings into zones (5), A* motion planning for the robot to move in-game items (6). Nodes are implemented in Python (except for the game interface, developed in QML) and inter-process communication relies on ROS. 6D poses are managed and exchanged via ROS TF.

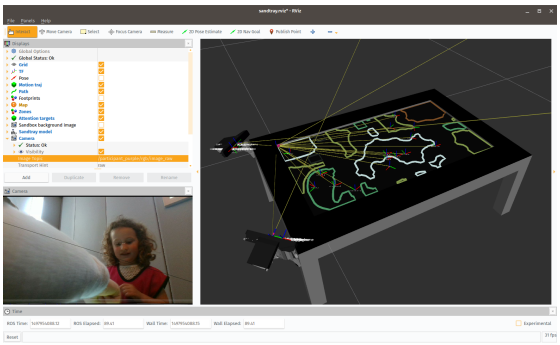


Figure 4: The free-play sandbox, viewed at runtime within ROS RViz. Simple computer vision is used to segment the background drawings into zones (visible on the right panel). The poses and bounding boxes of the interactive items are published as well, and turned into an occupancy map, used to plan the robot's arm motion.

the robot with more social inputs to test the sociability of a robotic controller (Fig. 3.3) and its impact on the interaction.

The robot's location is obtained by displaying fiducial markers on the touchscreen before the start of the interaction, so the transformation between the robot coordinate system and the touchscreen is known (Fig. 3.13). And this robot location can also be used to identify gazes from the child to the robot.

To make the children believe the robot is moving objects on the touchscreen, we synchronise a moving pointing gesture of the robot (Fig. 3.4) and a series of fake touches (Fig. 3.8) appied on the screen, moving the desired object. Once an object and a goal position have been selected, a planner (Fig. 3.6) generate a path for this image using the A* algorithm on an occupancy map obtained with the items footprints, then this plan is sent to a nodes synchronising the actuation on the robot and the fake touches on the game.

Other actions such as gaze, pointing or speech are also exposed as simple ROS topics.

Wizard-of-Oz To allow an experimenter to control the robot, a GUI to control the robot (Fig. 3.11) is provided and presents an identical representation of the state of the game on an other application which can be used on a tablet for example. The wizard can drag the objects in a similar fashion as what the child would do on the Sandtray, and on the release, the robot executes the dragging motion on the Sandtray, moving an object to a new location. The source code can be easily modified to add new specific buttons to execute other actions, such as having the robot talk to the child.

3.4 Experiment Manager

We have developed as well a dedicated, web-based, interface can be used by the experimenter to manage the whole experiment and data acquisition procedure (Fig. 3.10). This interface ensures that all the required software nodes are running, allow the experimenter to check the status and, if needed, to start/stop/restart any of them. It also help managing large data collection campaigns by providing a convenient web interface (usually used by the experimenter on a tablet) to record the demographics, resetting the game interface after each session, and automatically enforcing the acquisition protocol (see Table 1).

This interface has been extensively used to acquire the dataset that we present at Section 5.

4 Canonical procedures for data collection & analysis

The section presents *canonical* procedures to acquire data during testing, to pre-process it, and analyse it. We call them *canonical* because they are standard procedures, and where relevant, well integrated into the software pipeline of the sandbox (e.g., ROS integration) and represent state-of-the-art techniques. For the specific purpose of manually annotating the social interaction, we introduce as well a novel coding scheme, resulting from the synthesis of several existing techniques (Section 4.4 below).

However, these procedure are not normative. Researchers interested in reusing the free-play sandbox task for their own research would naturally adapt and extend these protocols to their own needs. Besides, certain aspects (most notably, the audio processing) are yet to be properly investigated.

Table 1: Data acquisition protocol

Greetings *(about 5 min)*
• explain the purpose of the study: showing robots how children play
• briefly present a Nao robot: the robot stands up, gives a short message, and sits down.
• place children on cushions
• complete demographics on the tablet
• remind the children that they can withdraw at anytime
Tutorial *(1-2 min)*
explain how to interact with the game, ensure the children are confident with the manipulation/drawing
Free-play task (up to 40 min)
• initial prompt: *"Just to remind you, you can use the animals or draw. Whatever you like. If you run out of ideas, there's also an ideas box. For example, the first one is a zoo. You could draw a zoo or tell a story. When you get bored or don't want to play anymore, just let me know."*
• let children play
• once they wish to stop, stop recording
Debriefing *(about 2 min)*
• answer possible questions from the children
• give small reward (e.g., stickers) as a thank you

4.1 Protocol

We typically adhere to the acquisition procedure described in Table 1 with all participants. To ease later identification, each child is also given a different and brightly coloured sports bib to wear.

Importantly, during the *Greetings* stage, we show the robot both moving and speaking (for instance, "Hello, I'm Nao. Today I'll be playing with you. Exciting!" while waving at the children). This is meant to set the children's expectations: they have seen that the robot can speak, move, and even behave in a social way.

Also, the game interface of the free-play sandbox offers a tutorial mode, used to ensure the children know how to manipulate items on a touchscreen and draw. In our experience, this has never been an issue for children.

Table 2: List of datastreams typically recorded. Each datastream is timestamped with a synchronised clock to facilitate later analysis.

Domain	**Type**
children	audio
	face (RGB + depth)
robot	full 3D pose
environment	RGB
touchscreen	background drawing (RGB)
	touches
	position and orientation of in-game items
static transforms between touchscreen and facial cameras	
cameras calibration informations	

4.2 Data collection

Table 2 lists the datastreams that are collected during the game. By relying on ROS for the data acquisition (and in particular the `rosbag` tool), we ensure all the ≈10 streams are synchronised, timestamped, and, where appropriate, come with calibration information (for the cameras mainly). In our experiments, cameras were configured to stream in qHD resolution (960×540 pixels) in an attempt to balance high enough resolution with tractable file size. It results in *bag* files weighting ≈1GB per minute.

In our own experiments, all the data (including up to 5 simultaneous video streams) was recorded on a single computer (quad core i7-3770T, 8GB RAM) equipped with a fast 4TB SSD drive. This computer was also running the game interface on its touch-enabled screen (sandtray), making the whole system compact and easy to deploy (one single device).

4.3 Data processing

Face and body pose analysis Off-line post-processing can be done on the images obtained from the cameras. We rely on the CMU OpenPose library Cao et al. (2017) to extract for both children the upper-body skeleton, 70 facial landmarks including the pupil position, as well as the hands' skeleton (when visible).

Further processing is possible: As the position of the camera, a potential robot and any object on the game is known, this landmarks can be mapped

to high level behaviours such as pointing or looking at an object. Additional analysis can be done on the facial landmarks to other social states, such as main emotion felt by the child.

Audio processing Similar processing can be applied on the audio stream. Library such as OpenSMILE provide audio features such as pitch and loudness contour, which inform on the general state of the child.

As of today, no reliable speech recognition engine exists for children Kennedy et al. (2017), but in the future, the audio should provide textual information on the requests and comments produced by the child.

Game interactions analysis Game features are also produced by the different nodes involved in the analysis of the game. The Playground segmentation produce a map of the regions based on the colour which can be used with the positions of the animal to identify from which zone to which zone an animal has been moved. The relative position of animal can also indicate if two animals have been moved closer. These relations and the drawing inform on what high level action the child is doing and can be used to infer the child's goal or desire.

4.4 Annotation of Social interactions

Annotating social interaction beyond surface behaviours is generally difficult. The observable, surface behaviours typically result of a superposition of the complex and non-observable underlying cognitive and emotional states. As such, these deeper socio-cognitive states can only be indirectly observed, and their labelling is typically error prone.

Our aim is to provide insights on the social dynamics, and we have synthesised a new coding scheme for social interactions that reuse and adapt established social scales. Our coding scheme (Figure 5) looks specifically at three axis: the level of *task engagement* (that distinguishes between *focused*, *task oriented* behaviours, and *disengaged* – yet sometimes highly social – behaviours); the level of social engagement (reusing Parten's stages of play, but at the micro-task level); the social attitude (that encode attitudes like *supportive*, *aggressive*, *dominant*, *annoyed*, etc.)

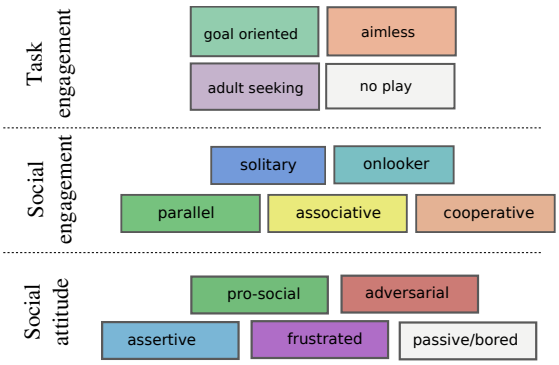


Figure 5: The coding scheme used for annotating social interactions occurring during free-play episodes. Three main axis are studied: task engagement, social engagement and social attitude.

Task engagement The first axis of our coding scheme aims at making a broad distinction between 'on-task' behaviours (even tough the free-play sandbox does not explicitly require the children to perform a specific task, they are still engaged in an underlying task: to play with the game) and 'off-task' behaviours. We call 'on-task' behaviours *goal oriented*: they encompass considered, planned actions (that might be social or not). *Aimless* behaviours (with respect to the task) encompass opposite behaviours: being silly, chatting about unrelated matters, having a good laugh, etc. These *Aimless* behaviours are in fact often highly social, and play an important role in establishing trust and cooperation between the peers. In that sense, they should not be discarded.

Social engagement: Parten's stages of play at micro-level In our scheme, we characterise *Social engagement* by building upon Parten's stages of play. These 5 stages of play are normally used to characterise rather long sequences (at least several minutes) of social interactions. Here, we apply them at the level of each of the micro-sequences of the interactions: one child is drawing and the other is observing is labelled as *solitary play* for the former child, *on-looker* behaviour for the later; the two children discuss what to do next: this sequence

is annotated as a *cooperative* behaviour; etc.

By suggesting such a fine-grained coding of social engagement, we enable proper analyses of the internal dynamics of a long sequence of social interaction.

Social attitude The constructs related to the social *attitude* of the children derive from the *Social Communication Coding System* (SCCS) proposed by Olswang et al. Olswang et al. (2006). The SCCS consists in 6 mutually exclusive constructs characterising social communication (*hostile*; *pro-social*; *assertive*; *passive*; *adult seeking*; *irrelevant*) and were specifically created to characterise children communication in a classroom setting.

We transpose these constructs from the communication domain to the general behavioural domain, keeping the *pro-social*, *hostile* (whose scope we broaden in *adversarial*), *assertive* (i.e., dominant), and *passive* constructs. In our scheme, the *adult seeking* and *irrelevant* constructs belong to Task Engagement axis.

Finally, we have added the construct *Frustrated* to describe children who are reluctant or refuse to engage in a specific phase of interaction because of a perceived lack of fairness or attention from their peer, or because they fail at achieving a particular task (like a drawing).

Video coding The coding is performed post-hoc with the help of a dedicated annotation tool (Fig. 6 which is part of the free-play sandbox toolbox. This tool can replay and randomly seek in the three video streams, synchronised with the recorded state of the game (including the drawings as they are created). An interactive timeline displaying the annotations is also displayed.

The annotation tool offers a remote interface for the annotator (made of large buttons, and visually similar to Figure 5) that is typically displayed on a tablet and allow the simultaneous coding of the behaviours of the two children. Usual video coding practices (double-coding of a portion of the dataset and calculation of an inter-judge agreement score) would have to be followed.

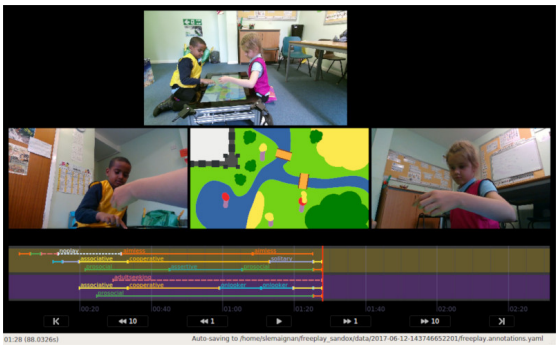


Figure 6: Screenshot of the dedicated tool developed for rapid annotation of the social interactions.

5 Baseline Datasets

We have been using the free-play sandbox task for an initial, large scale, data collection over a period of 3 months during Spring 2017.

This campaign aimed at (1) extensively evaluating the task itself (would children engage and exhibit a large range of social dynamics and behaviours?), (2) making sure the whole software architecture and data acquisition pipeline were reliable (they were), and (3) establishing two experimental baselines for the free-play sandbox task: the 'human' baseline on one hand (child-child condition), an 'asocial' baseline on the other hand (child - *non-social* robot condition). These two baselines are situated at the two ends of the spectrum of social interaction. They aim at characterising the qualitative and quantitative bounds of this social spectrum and can be used by the research community to evaluate given interaction policies.

A detailed description of the dataset is outside of the scope of this paper, and we only provide hereafter cursory informations on the dataset. Specific details regarding the methodology and the acquisition procedure can be found on the dataset website[3]. The dataset is open and accessible to any interested researcher, subject to adequate ethical clearance.

In total, 120 children were recorded for a total duration of 45 hours and 48 minutes of data collection. These 120 children (age 4 to 8) were split into two conditions: a child-child condition and a

[3] `https://freeplay-sandbox.github.io/`

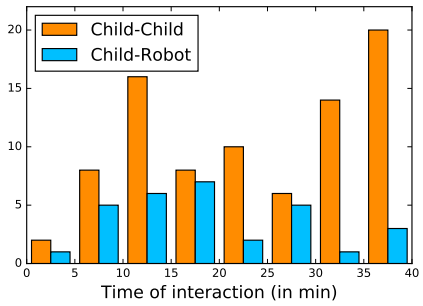


Figure 7: Durations of the interactions for the two conditions.

child-robot condition. In both condition, and after a short tutorial, the children were simply invited to freely play with the sandbox, for as long as they wished (with a cap at 40 min).

In the child-child condition (as seen in Figure 6), 45 free-play interactions (i.e., 90 children) were recorded with a duration M=24.15 min (SD=11.25 min).

In the child-robot condition, 30 children were recorded, M=19.18 min (SD=10 min). In this later condition, the robot behaviour was coded to be purposefully *asocial*: the robot would autonomously play with the game items, but would avoid any social interaction (no social gaze, no verbal interaction, no reaction to the child-initiated game actions).

Over the dataset, the children faces are detected on 98% of the images, which validates the location of the camera and the children to use the cameras to obtain facial social features.

Figure 7 presents an histogram of the durations of the interactions for the two baselines. The distribution of the child-child interaction durations shows that (1) all children engage easily and for non-trivial amounts of time with the task; (2) the task leads to a wide range of level of commitment, which is desirable: it supports the claim that the free-play sandbox is an effective paradigm to observe a range of different social behaviours; (3) long interactions (¿30 min) can result, which is especially desirable to study social dynamics.

In contrast, and notwithstanding the smaller number of participants, the distribution of the child-robot interaction durations shows these interactions are generally shorter. This is expected as the robot was explicitly programmed not to interact with the children, resulting in a rather boring (and at time, awkward) situation where the child and the robot where playing side-by-side – in some case for rather long periods of time – without interacting at all.

6 Discussion & Conclusion

6.1 Analysis of the free-play sandbox

The free play sandbox elicits a *loosely structured* form of play: the actual play situations are not known and might change several times during the interaction; the game actions, even though based on a single interaction modality (the touchscreen), are varied and unlimited (especially when considering the drawings); the social interactions between participants are multi-modal (speech, body postures, gestures, facial expressions, etc.) and unconstrained. This loose structure creates a fecund environment for children to express a range of complex, dynamics, natural social behaviours that are not tied to an overly constructed social situation.

The interaction is loosely structure. It is nonetheless structured: First, the physical bounds of the sandbox (an interactive table) limit the play area to a well defined and relatively small area. As a consequence, children are mostly static (they are sitting in front of the table) and their primary form of physical interaction is based on 2D manipulations on a screen.

Second, the game items themselves (visible in Figure 2) structure the game scenarios. They are iconic characters (animals or children) with strong semantics associated to them (like 'crocodiles like water and eat children'). The game background, with its recognizable zones, also elicit a particular type of games (like building a zoo or pretending we explore the savannah).

These elements of structure (along with other, less important, ones) make it possible for the free-play sandbox paradigm to retain some key properties that makes it a practical and effective scientific tool: because the game builds on simple and universal play mechanics (drawings, pretend play with characters), the paradigm is essentially

cross-cultural; because the sandbox is physically bounded and relatively small, it can be easily transported and practically deployed in a range of environments (schools, exhibitions, etc.); because the whole apparatus is well defined and relatively easy to duplicate (it essentially consists in one single touchscreen computer), the free-play sandbox facilitates replication of findings in HRI while preserving ecological validity.

6.2 Towards the machine learning of social interactions?

We presented a set-up and data set of relatively unconstrained interaction between children and between a robot and a child. The set-up captures a rich set of multimodal streams which can be used to mine the social, verbal and non-verbal communication between two parties engaging in a rich free-play interaction. The data holds considerable promise for training social signal interpretation software, such as engagement interpretation or eye gaze reading. The dataset collected has sufficiently rich data and a wide range of multi-modal dimensions making it particularly suitable for Deep Learning of social signal processing algorithms. It also allow for very rich input to action selection mechanisms needed for autonomous robot behaviour. Future work will focus on mining the data for social patterns occurring in play situations, as per Parten's classification, and will attempt to extract social signals relevant to drive the interaction. Some early results show, for instance, that deep learning shows considerable promise for high-resolution tracking of eye gaze from the RGB video streams.

Acknowledgments

This work has been supported by the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227).

References

P. Baxter, J. Kennedy, Senft E., S. Lemaignan, and T. Belpaeme. 2016. From Characterising Three Years of HRI to Methodology and Reporting Recommendations. In *Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference (alt.HRI)*. `https://doi.org/10.1109/HRI.2016.7451777`

Paul Baxter, Rachel Wood, and Tony Belpaeme. 2012. A touchscreen-based 'Sandtray'to facilitate, mediate and contextualise human-robot social interaction. In *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*. IEEE, 105–106.

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 82–90.

LB Olswang, L Svensson, TE Coggins, JS Beilinson, and AL Donaldson. 2006. Reliability issues and solutions for coding social communication performance in classroom settings. *Journal of Speech, Language & Hearing Research* 49, 5 (2006), 1058 – 1071.

Mildred B Parten. 1932. Social participation among pre-school children. *The Journal of Abnormal and Social Psychology* 27, 3 (1932), 243.

Laurel D Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012).

Int J of Soc Robotics (2017) 9:109–128
DOI 10.1007/s12369-016-0378-3

Nonverbal Immediacy as a Characterisation of Social Behaviour for Human–Robot Interaction

James Kennedy[1] · **Paul Baxter**[1] · **Tony Belpaeme**[1]

Accepted: 30 August 2016 / Published online: 16 September 2016

Abstract An increasing amount of research has started to explore the impact of robot social behaviour on the outcome of a goal for a human interaction partner, such as cognitive learning gains. However, it remains unclear from what principles the social behaviour for such robots should be derived. Human models are often used, but in this paper an alternative approach is proposed. First, the concept of nonverbal immediacy from the communication literature is introduced, with a focus on how it can provide a characterisation of social behaviour, and the subsequent outcomes of such behaviour. A literature review is conducted to explore the impact on learning of the social cues which form the nonverbal immediacy measure. This leads to the production of a series of guidelines for social robot behaviour. The resulting behaviour is evaluated in a more general context, where both children and adults judge the immediacy of humans and robots in a similar manner, and their recall of a short story is tested. Children recall more of the story when the robot is more immediate, which demonstrates an effect predicted by the literature. This study provides validation for the application of nonverbal immediacy to child–robot interaction. It is proposed that nonverbal immediacy measures could be used as a means of characterising robot social behaviour for human–robot interaction.

✉ James Kennedy
james.kennedy@plymouth.ac.uk

Paul Baxter
pbaxter@lincoln.ac.uk

Tony Belpaeme
tony.belpaeme@plymouth.ac.uk

[1] Centre for Robotics and Neural Systems, Cognition Institute, Plymouth University, Plymouth, UK

Keywords Nonverbal immediacy · Social behaviour · Robots for education · Social cues · Human–robot interaction

1 Introduction

Robot tutors are increasingly being explored as a means of delivering education to children in both dyadic [1–3] and larger group scenarios [4,5]. However, it remains unclear how a robot should behave socially in order to maximise learning outcomes. In the education literature, the social behaviour of a teacher is often assumed. For example, Kyriakides et al. [6] considers what makes teaching effective and lists how lessons are structured, how learning is assessed, how time is managed, and so on. The role of social behaviour is not mentioned; we believe that this is because it is so fundamental that it is assumed to be present. A base level of sociality can reasonably be expected when interactions occur between humans, but when the tutor is a robot, this element becomes unknown. The fundamental assumption of social behaviour for teaching highlights it as an important element to resolve.

Various researchers have begun to address certain aspects of social behaviour for educational contexts in human–robot interaction (HRI). Gordon et al. consider the impact that the curiosity of a robot may have on reciprocal curiosity of a child and their subsequent learning of words. The human–human interaction (HHI) literature predicts an increase in learning as curiosity increases, however this finding was not replicated with robots [1]. Saerbeck et al. also consider language learning with a socially supportive robot, where the socially supportive robot leads to more retention than a robot without this social behaviour [7].

Personalisation of interactions has been explored in health education for children with diabetes. In a dyadic interaction with a robot, the robot would ask the child for various items of personal information (name, favourite sports and favourite colours) and use them during the interaction [8]. The personalised robot provided an indication that children's perceived enjoyment of learning was enhanced, although too few subjects took part to make conclusions about learning effects. Other authors have personalised human–robot interaction in learning contexts through manipulating the timing of lessons [9], or through setting personalised goals [10]. However, this becomes more about teaching strategy and does not help to generate lower-level social behaviour.

Personalisation has also been incorporated into larger scale social behaviour changes in interactions where children learn about prime numbers [2]. A surprising result was found where a robot designed to be 'more social' did not lead to learning gains, whereas children interacting with a 'less social' robot did experience significant learning gains. Such labelling raises questions about how HRI should characterise sociality: what constitutes being more or less social, and how can this be measured and expressed in experimental reports? This is an important issue to resolve to ease the understanding and interpretation of results, and for comparisons to be made between studies, often in differing contexts.

This paper seeks to explore one way in which sociality might be characterised for HRI: *nonverbal immediacy*. The elements of nonverbal immediacy are broken down into individual cues (such as gaze, gesture, and so on) and considered for use in an educational context, before being brought back together into an implemented behaviour to evaluate whether the concepts hold true in practice with robots. The rest of this paper is structured as follows. First, the social context of learning and the concept of nonverbal immediacy are introduced (Sect. 2). Nonverbal immediacy will then be considered in terms of the component social cues by which it is measured; the effect of each social cue on learning will be explored from both a HRI and a HHI perspective (Sect. 3). This will culminate in a set of guidelines for robot social behaviour during educational interactions (Sect. 5). These guidelines are used as a basis for an evaluation in which nonverbal immediacy is measured and compared to recall. The study uses a 2×3 design, comparing nonverbal immediacy scores and recall between children and adults, depending on whether they have seen a high immediacy robot, a low immediacy robot or a human reading a short story (Sect. 6). A discussion of the potential benefits and limitations of this approach will be carried out (Sect. 7), with the suggestion that nonverbal immediacy is a useful means of characterising and devising social behaviour for robot tutors.

2 Sociality, Immediacy and Learning

It has long been posited that the role of society and social signals are of great importance in teaching and learning, most notably in Bandura's Social Learning Theory [11] and Vygotsky's Social Development Theory [12]. The importance of social signals is apparent from a young age, with social cues playing a role in guiding attention and learning [13]. However, we still have relatively little understanding of what impact combinations of multimodal social cues have on learning in complex settings [14]. Correspondingly, we don't seem to be able to correctly identify highly effective teaching when we see it, raising questions about how to define what effective teaching consists of [15].

Social interaction can be considered as the bond between cognitive processes and socio-emotional processes [16]. The outcome of such interaction can be measured through social performance or learning performance, either of which can in turn reinforce the cognitive or socio-emotional processes taking place in an individual (Fig. 1). This concept is supported through definitions of learning, which can be broken down into 'affective' and 'cognitive' learning [17]. Social interaction has the ability to influence both of these learning elements, and indeed HRI researchers have sought to do just this. Some researchers have focussed on the social behaviour of the robot with the aim of influencing cognitive processes [18], whereas others have sought to influence the socio-emotional processes to a greater extent [19].

Many studies considering the impact of social behaviour use a human expert or model in order to inform the behavioural design for a largely autonomous robot, for example [2,20]. Additionally, many studies only vary a limited set of social cues, often to tightly control the experimental conditions [21–23]. Whilst these approaches allow us to learn about the impact of some social behaviour on learning, there are many difficulties in comparing between studies as there

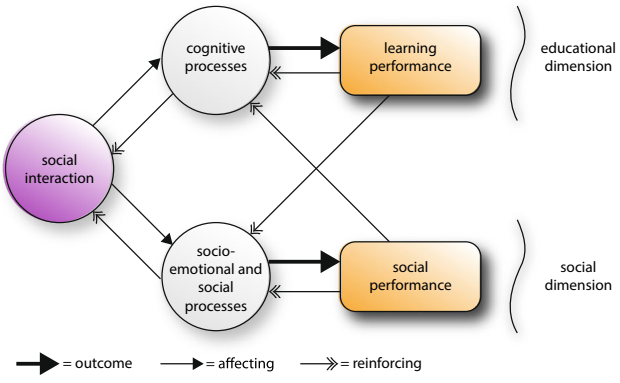


Fig. 1 A depiction of the role of social interaction for an individual, with two possible outcomes: social performance and learning performance—adapted from Kreijns et al. [16]

is no common metric for the overall social behaviour of the robot. It is also unclear what would happen when multiple social cues are modified together; it seems plausible that the effects found from single cue manipulation would be additive, but there is evidence to suggest that humans do not process social cues in this manner [24]. A means of characterising social behaviour across multiple contexts would therefore provide a great advantage to the field for making cross-study comparisons.

One possible concept for making such social characterisations is *nonverbal immediacy*. Immediacy can be defined as "the extent to which communication behaviours enhance closeness to and nonverbal interaction with another" [25], with closeness referring to both proximity and psychological distancing. Nonverbal immediacy is a measure of nonverbal behaviour which indicates a "communicator's attitude toward, status relative to, and responsiveness to" an addressee [25]. Richmond et al. [26] developed a highly reliable questionnaire to measure nonverbal immediacy in communication contexts. The 'Nonverbal Immediacy Scale-Observer Report' developed is freely available online[1] and incorporates the following social cues into a single measure: gestures, gaze, vocal prosody, facial expressions, body orientation, proximity, and touch.

Nonverbal immediacy emphasises the multimodal nature of interaction and the consideration of all social cues taken in context with respect to each other. The measure provides a characterisation of 'sociality' which can then be correlated against an outcome, such as learning, and compared against another set of behaviour characterised in the same manner. It has found extensive application in educational research, most often in university lecture scenarios [27].

When reviewing the literature surrounding nonverbal immediacy it is important to make the distinction between 'affective learning', 'cognitive learning' and 'perceived cognitive learning'. *Affective* learning considers constructs such as attitudes, values and motivation towards learning [28]. *Cognitive* learning typically focusses on topic specific knowledge and skills [29]. *Perceived* learning is a measure of how much students believe they have learnt, or how confident they are in what they have learnt, such as in [30]. Whilst the correlation with measured cognitive learning gains is only moderate, relatively few studies have used experimental measures; most have used perceived learning, which has a particularly strong correlation with teacher immediacy [27]. It has been experimentally found that perceived learning and actual recall are moderately correlated in such contexts [31], so whilst perceived learning is not as strong as measuring actual learning, it can at least be used as an indication of the nature of relationships.

[1] http://www.jamescmccroskey.com/measures/nis_o.htm.

A positive correlation between nonverbal immediacy and perceived cognitive learning has been validated across several cultures, including the United States, Puerto Rico, Finland and Australia [32]. From this McCroskey et al. postulate that expectation of immediacy plays a key role in how cues are interpreted, presenting opportunities for high immediacy teaching to have a strong positive impact in generally low immediacy cultures, but a negative impact for low immediacy teaching in high immediacy cultures [32]. A similar suggestion relating to the use of robot social cues in teaching contexts has also been raised in HRI [33].

Both verbal and nonverbal immediacy behaviours have been shown to lead to an increase in motivation, and, in turn, student learning [34,35]. In some cases, such as in a task to recall contents of a lecture [36], cognitive learning gains are not found, but affect for the instructor and material increases when the instructor is more nonverbally immediate. However, there are other examples demonstrating a link between greater nonverbal immediacy and increased recall [37,38]. A more extensive review of the potential benefits of immediacy (both verbal and nonverbal) can be seen in [39].

Nonverbal immediacy has been studied only briefly in HRI contexts before. Szafir and Multu [23] use it as a means of motivating and evaluating robot behaviour during a recall task with adults. In line with literature studying nonverbal immediacy with humans, they find that as immediacy increases, so does recall. The adults were also able to notice when the nonverbal immediacy of the robot had increased, confirming that people are sensitive to such cues in robots. Nonverbal immediacy concepts have also been used by the same lab to motivate behavioural manipulations for persuasive robots [40]. However, it should be noted that it doesn't appear that a complete nonverbal immediacy questionnaire was used in either of the studies. This is important as it is argued in this paper that a key motivator for using nonverbal immediacy measures is the consideration of all cues taken in context; this idea will be returned to and expanded upon in Sects. 4 and 7. Finally, nonverbal immediacy has recently been proposed for use in HRI studies to motivate exploring the perception of a robot when posture and nodding behaviour is varied [41].

3 Social Cues of Nonverbal Immediacy

Based on the method used to calculate nonverbal immediacy, if there is a linear relationship between learning and immediacy (as suggested by [34]) then learning would be maximised if the social cues used in nonverbal immediacy are maximised. However, there are also suggestions that the relationship may not be wholly linear in nature [42,43]. As such, it remains slightly unclear how immediacy should be utilised for social robots. The following subsections will con-

sider each of the component cues which form the nonverbal immediacy measure in turn to provide further insight into how they can be applied in practice, with a particular focus on findings from HHI and HRI learning scenarios. The aim is to generate guidelines for social behaviour in robot tutoring scenarios that are informed by the concepts of the nonverbal immediacy measure and supported by previous work in both HHI and HRI (Sect. 5).

3.1 Gestures

Gestures play an important role in teaching and learning [44,45]. Children are more likely to repeat the speech of a teacher if a matching gesture accompanies the speech when compared to the same speech without a gesture, but less likely with a mis-matched gesture compared to no gesture [46,47]. This basic recall is a first step towards learning. Furthermore, these studies show that children can use gestures in understanding problem-solving strategies, giving them the potential to learn both through problem solving and how to approach solving problems.

For young children, it has been suggested that gesture use (specifically symbolic gestures) can facilitate cognition [48]; possibly because gestures can lighten cognitive load, lending more resources to memory tasks [49]. Indeed when children are slightly older (aged 8–10) gestures can help learning to 'last' for longer, with correct answers in an algebra follow-up test four weeks after a learning session staying higher in a gesture and speech condition than in a speech only condition [50]. Equally, gestures made by children can be used to assess their learning [51], with adults able to be more certain of their judgements of children's learning when their gestures matched their verbal explanation.

Such findings are reinforced in studies concerning instructional communication for learning, with children's performance improving more when given instructions with gestures as opposed to without in a symmetry recognition test [52]. These findings seem to have been partially replicated in HRI, with a robot utilising contingent gesturing leading to increased recall of material from a presentation [23]. However, precisely how to use gestures to influence learning in HHI is an open field with many questions still necessitating futher exploration [53]; this is even more true for HRI where less work examining the use of gesture and learning has been conducted.

The use of hands seems to be particularly important. It is not just the orienting of attention, such as with a laser pointer, but the fact that the gesture is done with a hand that leads to an improvement in learning [54]. It has been shown that humans can accurately interpret pointing by a humanoid robot (an Aldebaran NAO), but that for best results, the arm on the side which the object to be pointed at should be used [55]. However, whether the hand of robot has the same attentional and learning impact as that of a human is not known. It has also been established that being present (as opposed to on video) does not affect how much attention gestures draw between humans [56], but no such study comparing humans and robots could be found.

3.2 Gaze

From an early age, children use social cues such as eye gaze to help direct their learning. Despite social cues distracting briefly from the material to be learnt, infants learn more with gaze cues present than when their learning is not directed by such cues [57]. These positive effects have also been successfully implemented in computational models [58]. Even at 15 months old, children have a tendency to use the gaze of a social interaction partner, instead of distracting and erroneous saliency cues for word learning associations [59]. The power of gaze, or even just the eyes, in influencing behaviour is still observed in adults, with surprisingly strong results. For example, just an image of eyes near a donation point can increase charitable donations by almost 50 % [60].

Selective processing of social cues for learning has far-reaching implications for human–robot interaction. Head movement alongside eye gaze can assist humans in responding to robot cues [61]; use of this social cue could have advantages in learning. However, this has not been found in infants learning from robots, where they follow the gaze direction of both a robot and a human, but only the human gaze facilitated the learning of an object [62]. It was suggested that this could signify a disposition of infants to consider humans a superior source for learning. It remains to be seen whether this holds true for slightly older children, or with children more familiar with the concept of robots. Equally, this result could be a demonstration that humans process robot gaze in a cognitively different manner, as argued in [63].

College students who receive gaze at the start of each sentence when receiving verbal information can recall significantly more than those who receive no gaze [64]. This holds true for both simple and difficult material, for both genders. It is hypothesised that this is because the interaction feels more 'intimate' and prevents mind-wandering whilst receiving the information. These findings have also been shown to occur with younger children, aged between 6 and 7 [65]. Greater gaze from a storyteller led to increased recall from children when subsequently asked questions, compared to those in a lesser (but still some) gaze condition. This study reveals a trend towards possible interaction effects between the information content, gender and gaze, speculating that females are less affected by gaze than males when the material is more difficult.

Logically, it follows that using appropriate robot gaze towards a child might be beneficial for recall and learning.

Work done in virtual environments demonstrates that caution must be used, as simply staring at a human interactant actually reduces their willingness to engage in mutual gaze, despite increased opportunity [66]. It should be noted that this difference in mutual gaze did not actually translate to a difference in task performance, but this was hypothesised as being due to the relative simplicity of the task. A lack of effect due to gaze has been observed in human–robot interaction studies as well. In both [67] and [68], a tutoring robot received more gaze from children, which could theoretically be beneficial for child learning (as the robot is delivering learning content), but no learning differences were found.

Nevertheless, the outcome here is a message of balance: gaze can clearly have positive effects on learning [58,62,64,65], but if it is not meaningful, or is too abundant then it can discourage mutual gaze, thereby limiting potentially positive effects [66]. This remains a challenge, as it is not trivial to decide how much gaze is 'just right', or precisely when a gaze should be made by a robot.

3.3 Vocal Intonation/Prosody

The voice that an agent uses can dictate how much they are liked and how hard humans try to understand the material they are presented with [69]. Those who interacted with an agent who had a human voice preferred the agent and also did better in learning transfer tests when compared to those who interacted with the same agent with a machine-synthesised voice. The sound of a voice can have a significant impact on retention and transfer of a novel subject when presented through narration [70]. Retention is better when a voice has a 'standard' (as opposed to foreign) accent and is human rather than machine-like, as well as being more likeable in both cases.

However, this result was found with college students and virtual agents. It has not been established whether this effect is also observed outside of this restricted demographic, nor whether specific embodiments of robots create expectations that violate these rules. For example, it may be less appropriate to have a deep male human voice when using a robot such as the Aldebaran Nao[2] than a RoboThespian.[3] It is suggested that a possible uncanny valley effect [71] may occur, where participant expectations are violated when a human voice is played alongside a not-convincing-enough animated agent. An indication in this direction has been found with virtual agents, where participants preferred an animated agent with a machine-like voice and a non-animated agent with a human voice [72].

Vocal intensity can also be used to influence learning. Compliance, a factor in learning, can be increased through raising vocal intensity, as in [73]. This HHI study was conducted in a public space where compliance was greatest when using a medium level of vocal intensity; around 70 dB. It is likely that this level would need adjusting depending on the ambient noise in the space a robot tutor would be acting in, and how far from a student it would be. Vocal intensity has successfully been combined with gestures in a model which is based on nonverbal immediacy to improve attention and recall of a human in an HRI presentation scenario [23]. Whilst not confirming all of the results discussed in this section relating to vocal prosody, it certainly demonstrates that there is great potential for many of the same principles from HHI being applied to HRI with positive results.

Interestingly, speech rate appears to have a significant impact upon perceptions of nonverbal immediacy, but not on recall [74]. As speech rate increases, perceived immediacy of a speaker goes up, but there is no significant difference in recall as a change of immediacy might predict. This could potentially be explained by the capacity of humans for speech. The average human speech rate is 125–150 words per minute, but learners have twice as much cognitive capacity, being able to process speech at 250–300 words per minute [75]. This gives great scope for increasing speech rate, and therefore immediacy, but without any great change in terms of the listener's cognitive processing.

3.4 Facial Expression

In a HHI study examining the relationship between the social cue elements of nonverbal immediacy and cognitive learning across a number of different cultures it was found that alongside gaze and vocal prosody, smiling from the teacher was one of the more strongly correlated cues to student learning [32]. This result has also been replicated more recently [76], additionally showing the positive relationship between nonverbal immediacy and motivation (with facial expressions having a large effect size).

Experimental data from human–computer interaction (HCI) with an embodied conversational agent revealed no significant difference in recall of subjects when interacting with an agent which was either neutral, or able to express joy and anger [77]. Several reasons are put forward as to why this may have been the case, including a ceiling effect within the task, the amount each emotion was displayed, or that the facial expressions were simply ignored in favour of focussing on the task. As such, it is unclear whether the benefits of facial expression seen in HHI will translate to HCI and HRI.

Despite the suggested impact of facial expressions on learning or motivation in HHI, no data could be found regarding the impact of learning and facial expressions of robots. A possible explanation is that much of the research to-date regarding learning in HRI is performed with robots

[2] https://www.aldebaran.com/en/humanoid-robot/nao-robot.

[3] https://www.engineeredarts.co.uk/robothespian/.

such as the Aldebaran NAO, Keepon, and Wakamaru which have largely non-manipulable faces. Due to the movement required in expressing facial emotion, the uncanny valley [71] could also be a current limitation for robots.

3.5 Proximity and Body Orientation

The proximity between interactants is correlated to compliance effects [78]. It is suggested that a distance of 1–2 feet (30–60cm) is optimally conducive to compliance between humans (from studies conducted in Western cultures) [79], however whether this is the same for HRI has not been established. This is possibly because judging the physical proximity at which a robot should be from a student would not necessarily be as simple as a strict 1–2 feet (30–60 cm) rule. In human interactions, verbal feedback can modulate (positively and negatively) the proxemic impact on compliance [80]. In HRI, comfortable distances are dictated through the complex interplay of factors such as the size of the robot [81], how much the robot gazes towards a human and how likeable they previously perceive the robot to be [82].

Only about 60 % of people conform to the same proxemic social norms with robots as they do with people [83]. That being said, compliance effects have been seen in educational interactions between children and robots at a distance of about 2 feet (60 cm), although this hasn't been compared against a control with closer or further distances [84]. Additionally, it would appear that younger children have a smaller personal space, presumably due to their smaller size, so further work would need to be done for people of different sizes [85].

Research conducted with a robot in a variety of task contexts show humans generally prefer the robot to be 0.46–1.22 m away [86]. However, it is warned that the dynamic nature of interaction with a robot should not necessarily be reduced to a simplistic rule. Indeed, the previous paragraph suggested the impact of variable robot appearance and behaviour, but there are also environmental and task factors to consider. For instance, if it is important to hear speech in a noisy environment, then it might be that a closer distance between interaction partners is more comfortable, when outside of these parameters it would usually not be.

Several design guidelines for robotic proximity are presented in [87]. It is suggested that people who are familiar to the robot can be approached more closely, to direct gaze away from the face of a human as an approach is made, and to factor in the human's attitude towards robots when maintaining distance. The impact of human attitude towards robots is further supported experimentally in [88] where the necessity of building rapport before increasing closeness is emphasised. This could be an important factor in teaching in order to gain compliance.

Studies directly examining the impact of body orientation on learning could not be found; this is possibly due to the entanglement of body orientation with many other social cues. If not orientated to an interaction partner only limited eye gaze will be possible, gestures may be occluded and it may be more difficult to hear any speech. Nor could any studies be found studying the specific impact of co-located physical proximity on learning; most work considers co-located learning against distance learning (not co-located), but this then becomes about social presence rather than proxemics. Logically, it would seem reasonable that a middle-ground should be sought. The robot should not be too far away as then the student may struggle to perceive verbal instructions and nonverbal signals. If more compliance is required, then a closer distance should be sought. Further research is required to decide what is to be considered 'too close' in specific scenarios, with humans of certain ages and certain robot sizes/designs; work such as [83,89] provides a strong starting point in this direction.

3.6 Touch

Touch has been shown to lead to a positive affective state in HHI, even with very short touches and when subjects were unaware of the touch [90]. This positive response to touch has also been shown in HRI. When a robot offered an 'unfair proposal' to participants with touch, their EEG response showed less negativity towards the robot than when the robot did not touch as they made the proposal [91]. Of course, liking does not necessarily result in better learning, but there are indications that if students like an instructor more they will achieve more highly [92].

Touch has also been linked with compliance [93], a useful tool for teachers when they need to influence students in order to get them to engage with lessons. The potential for utilising touch in HRI and educational contexts has previously been highlighted [94] but, as yet, remains underexplored.

4 Synchrony and Multimodal Behaviour

Of course, social cues do not occur in isolation, neither from other cues, nor from the environment and the interaction they are being used in. Behaviour is multimodal, and the cues must be contingent with respect to the interaction and congruent with other social cues being utilised in order to be interpreted correctly and efficiently. Social cues could be perceived as a single percept, which requires that cues be considered as an integrated whole [24]. Nonverbal immediacy is measured by taking many social cues into consideration with respect to one another, and thus supports the principles behind interpreting social behaviour in this manner.

Table 1 Behavioural guidelines for robots in educational contexts derived from the nonverbal immediacy and social cue literature

	Guidline	Caveat (if applicable)	Section Ref.
G1	In general, mutual gaze should be sought as more mutual gaze leads to increased recall	A robot should not fixate its gaze at a human for prolonged periods of time or they will avoid mutual gaze	2.2
G2	HCI suggests that vocal intonation/prosody should be of the same accent as the participant and human-like rather than machine-like	This remains under-explored in HRI	2.3
G3	For best compliance, vocal volume should be 70 dB in public spaces	Adaptivity to ambient noise may be required depending on the scenario	2.3
G4	Gestures should be relevant to verbal content being delivered and should be used to aid understanding		2.1
G5	Use of hands (as opposed to laser pointers, or similar) is key in directing learner attention		2.1
G6	When using pointing to direct attention, it is important to use the arm on the same side as the object being pointed to		2.1
G7	Closer proximity should be sought for increased compliance. For humans a guideline is around 1–2 feet (30–60 cm)	Appropriate distances for robots are not well established and could depend on the size of the robot	2.5
G8	Nonverbal Immediacy measures suggest that a relaxed body position, leaning forwards, is more immediate (and therefore leads to increased learning gains)		2.5

These concepts are exemplified experimentally by Byrd et al. [95] who further explored the conclusions drawn from studies such as those done by Cook et al. [50] regarding gestures and learning (discussed previously in Sect. 3.1). They found that when children did not copy eye movements accompanying gestures the lasting learning effect disappears.

Support for the role of synchrony in social cues can be seen in [96,97]. Head gaze, gestures and spoken words were all used to direct attention. When any of the cues were incongruent (e.g. responses had to be made to head-gazes, whilst a pointing gesture was made in a different direction), interference effects were found, slowing down responses. If social cues are not synchronous and congruent then interactions will likely be impeded by this additional processing time.

Not just the cues being used, but also their contingency can influence interactions. A robot which displays more contingent social cues, such as appropriate gaze and pointing gestures, can elicit greater participation in an interaction [98]. When applied to an educational context, it is reasonable to suggest that greater participation will lead to an increase in learning [99].

5 Guidelines

Based on the analysis of the individual cues that comprise nonverbal immediacy (Sect. 2) we seek to derive a set of design guidelines that can be applied to HRI in tutoring contexts. Nonverbal immediacy and learning have been positively correlated in human–human studies, and there have been indications that this may be supported in HRI as well [23]. The social cues which make up nonverbal immediacy have been explored through the HHI and HRI literature, often revealing a connection with learning gains on an individual basis, providing some insights into the practical application of such cues for HRI. From this, guidelines for robot social behaviour in educational interactions have been devised (Table 1).

6 Evaluation

If an effect seen in HHI studies concerning nonverbal immediacy can be replicated with robots, then this strengthens the case for phenomena correlated with immediacy in HHI studies transferring to HRI as well. This could provide useful links to a body of literature from which insights into design of robot behaviour could be derived.

The guidelines in the previous section use nonverbal immediacy as a basis for behaviour generation, which is commonly measured through observational reports, such as those seen in [26]. This measure has seen limited application in HRI evaluations before, though where it has, the immediacy scores have not been explicitly stated [23,40]. As such, it would be beneficial to validate that behaviour intentionally created as more or less immediate is judged as such when applied to robots, as it is with humans. Additional validation with children (due to the educational context of this work)

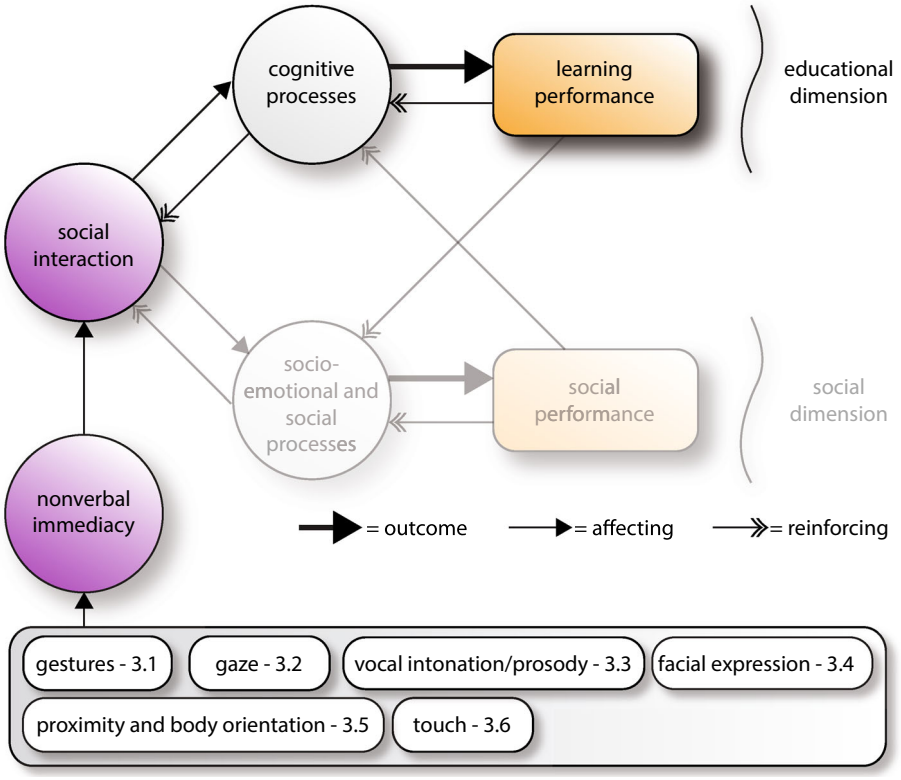


Fig. 2 Updated version of Fig. 1 depicting the influence of nonverbal immediacy on social interaction, and the educational dimension of social interaction which this paper is concerned with. Section references are provided in the diagram for each of the social cues that nonverbal immediacy consists of

to check whether they interpret the behaviour in the same manner as adults would allow the guidelines to be applied to a larger range of HRI scenarios. A human condition is therefore used to provide a reference point for the child ratings with respect to the adult ratings. This will enable an assessment of the reliability of child ratings of immediacy (which does not readily appear in the literature), as a basis for the subsequent examination of child ratings of robot immediacy. The comparison between child and adult interpretation of human nonverbal immediacy serves as a useful intermediary step between the existing literature and applications of nonverbal immediacy with robots and children. The evaluation here focuses on the outcome of the educational dimension of social interaction (as opposed to the social dimension) as influenced by nonverbal immediacy (Fig. 2).

6.1 Methodology

A 2 × 3 condition study was devised to explore how nonverbal immediacy would impact recall; two factors which have been shown to be positively correlated (Sect. 2). In order to evaluate whether children and adults interpret the behaviour of a robot and a human in the same way, a scenario which could be understood by both groups was required. As such, the study design started from the perspective of the children (who are presumed to have a shorter attention span and more limited knowledge in some areas such as vocabulary) and was then applied to adults. Recall of a presented short story was decided to be an appropriate task for this purpose as this matched the methodologies of immediacy studies.

Participants A total of 117 participants took part in the study, but one child had to be excluded due to an incomplete questionnaire and two adults were excluded due to inconsistent online video timestamps; this will be expanded on later in this section. 83 children (age M = 7.8 years, SD = 0.7; 47 F, 36 M) and 31 adults (age M = 23.5 years, SD = 3.9; 7 F, 24 M) remained for data analysis. All participants consented to participation in the study and all children had parental permission to take part. The children were recruited from one school year group of a primary school in the UK; the children were split across conditions based on their usual school classes, which ensures an appropriate balance for gender and academic ability. Adults in the robot conditions were recruited through regular lectures, and through online advertising for the human condition.

Short Story A short story was created for the purpose of the recall test. The story was largely based on one freely available from a website containing many short stories for children.[4] This was done to make sure that the language and content was appropriate for children. Some elements were added or modified in order to create opportunities for recall questions, and some of the phrasing was modified so that the robot text-

[4] http://freestoriesforkids.com/children/stories-and-tales/robot-virus.

to-speech sounded more accurate. The final version of the story created can be seen in Appendix 1 and lasts for just under 4 min when read in the experimental conditions. None of the participants reported to have heard or read the story before.

Measures Two measures were used: a nonverbal immediacy observer report questionnaire and a recall test. The Robot Nonverbal Immediacy Questionnaire (RNIQ; Appendix 2) was based on the short form of the Nonverbal Immediacy Scale, sourced from [100] and freely available online.[5] Exactly the same questionnaire was given to both children and adults. The questionnaire was modified from the original to make it easier to understand and complete for children. This was done in four ways:

1. "He/she" was changed to "The robot", or "The man" depending on the condition.
2. "while talking to people" was changed to "while talking to you".
3. The response of 'occasionally' was changed to 'sometimes'.
4. Instead of filling in a number at the start of each line, boxes labelled with the scale were presented for each question. This prevents children from having to keep referring back to the top of the page and potentially losing their thought process, and also prevents mistakes in interpreting their handwriting during analysis.

The recall test was devised based on information provided in the short story and consisted of 10 multiple choice questions, with a final free text answer about the moral of the story. The full list of questions and answer options can be seen in Appendix 3. The questions were designed to vary in difficulty based on how many times the piece of information had been stated, how central it was to the plot, and how many answer options were similar to the correct one. An additional question was added to the adult human condition regarding the colour of the background in the video; this was part of a series of checks to ensure that the video had actually been watched.

Hypotheses and Conditions Based on the literature explored in Sect. 2 and the guidelines in Sect. 5, four hypotheses for the study were considered:

- **H1**: Robot behaviour designed to be more or less immediate will be perceived as such, as measured through the nonverbal immediacy scale.
- **H2**: Children and adults will perceive nonverbal immediacy in the same manner for both robots and humans (i.e. children and adults ranking of immediacy will agree).
- **H3**: Recall of the story will be greater when read by a character with higher nonverbal immediacy.
- **H4**: As nonverbal immediacy of the character reading the story is perceived to increase by an individual, their recall of the story will also increase.

In order to address these hypotheses, three conditions were devised which were shown to both children and adults:

1. **High nonverbal immediacy robot** (Fig. 3 *centre*)—using the guidelines in Sect. 5, the robot behaviour was maximised for immediacy where possible; full details of the robot behaviour can be seen in the following paragraph. Child $n = 27$; adult $n = 9$.
2. **Low nonverbal immediacy robot** (Fig. 3 *left*)—using the guidelines in Sect. 5, the robot behaviour was minimised for immediacy where possible; full details of the robot behaviour can be seen in the following paragraph. Child $n = 28$; adult $n = 9$.
3. **Human** (Fig 3 *right*)—a human was recorded on video reading the story. This was to ensure identical behaviour between child and adult conditions and to time the story to be at the same pace as the robot conditions in order to have equivalent exposure time and reading speeds (which can impact recall [74,101]). This condition enables the immediacy ratings of children to be validated with respect to adults. The human was not given explicit instructions in terms of nonverbal behaviour, as their immediacy level is not under consideration, but whether the children and adults perceive their immediacy level in the same way is. Therefore, the behaviour itself is not of concern, provided that it is identical between conditions (the video recording ensures that this is the case). Child $n = 28$; adult $n = 13$.

Robot Behaviour The high and low nonverbal immediacy robot conditions were developed based on the guidelines from Sect. 5. The conditions sought to maximise the differences between the behavioural dimensions which the guidelines address (and therefore also the dimensions measured by the nonverbal immediacy scale). Some dimensions were not varied due to limitations in the experimental set-up. Facial expressions were not varied as the robot being used for the study, an Aldebaran NAO, is not capable of producing facial expressions such as frowning or smiling. Proximity was not varied due to the group setting in which the study was being conducted. When the robot is telling the story to a classroom of children it is not feasible, or safe, to incorporate touch or to approach the children. The operationalization of behavioural manipulations that were carried out can be seen in Table 2.

Procedure For the robot conditions, the robot was placed at the front of the classroom on a table to be roughly at

[5] http://www.jamescmccroskey.com/measures/nisf_srni.htm.

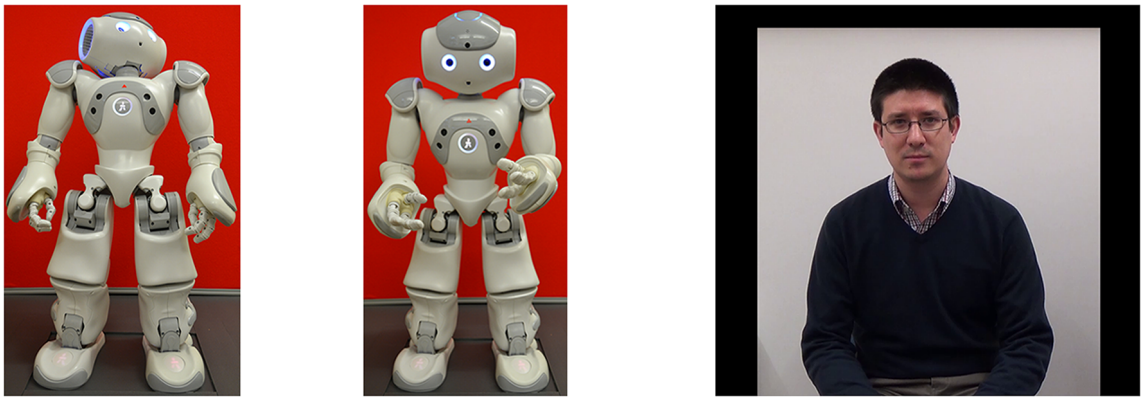


Fig. 3 Still images from the conditions used in the evaluation; *left* to *right*: (1) low nonverbal immediacy robot, (2) high nonverbal immediacy robot, (3) human. *Red backgrounds* for the robot were not used in practice and are just used to ease visibility here; the video was shown in widescreen format, with a *black background* covering the unused space, as in the figure

Table 2 Operationalization of behavioural manipulations between robot immediacy conditions

Behavioural dimension	High nonverbal immediacy	Low nonverbal immediacy
Gesture	Frequent gestures, occurring approximately every 12 seconds during the story. Slight randomness added to joints to provide small constant movement	No gestures, no joint random movement
Gaze	Head gaze directed forwards randomly at approximately the same height as the robot towards the centre of the movement range (towards observers)	Head gaze directed randomly up and towards the corners of movement range (over/away from observers)
Vocal prosody	No modifications to standard text-to-speech (TTS) engine, allowing shaping of sentences and responsiveness to punctuation	All strings passed to TTS have punctuation stripped and are forced to be spoken with no context of the sentence (resulting in words sounding identical every time they are said). Additionally, vocal shaping was reduced via a TTS parameter
Body orientation	Leans towards observers by approximately 15 degrees	Leans away from observers by approximately 15 degrees

the head height of observers (either children or adults). The experimenter would then explain that the robot would read a story and that afterwards they would be required to fill in a questionnaire about what they thought of the robot. The recall test was explicitly not mentioned to prevent participants from actively trying to memorise the story. The experimenter then pressed a button on the robot's head to start the story. Once the story was complete, the nonverbal immediacy questionnaires were provided to all participants. When the whole group had completed this questionnaire, the recall test was introduced and given to participants. For the children, this was followed by a short demonstration of the robot. The human video condition procedure was the same for the children. The video was resized to match the size of the robot as closely as possible, and the volume was adjusted to be approximately the same as well.

As the children did not know this person, the adults should not either so that the reported immediacy score is based purely on the behaviour seen in the video and not prior interaction. The subjects for the video condition were recruited online and completed a custom web form which prevented the video from being paused or played more than once, and recorded timestamps for the start of the video, the end of the video, and the completion of the questions. An additional question was also added to the recall test to verify that the participants had actually watched the video (as opposed to the rest of the recall questions which can be answered through listening alone). One participant was excluded from analysis as the timestamps for the start and end of the video indicated too little time for the full video to have been viewed and another participant was excluded as the time between watching the video and completing the questions was in the order of hours (all other participants completed all questions in under 10 min), indicating that the intended protocol had been violated.

Table 3 Mean nonverbal immediacy scores by condition

Condition	Adult *M*	95 % CI	Child *M*	95 % CI
High immediacy robot	50.2	[47.0, 53.5]	50.8	[48.6, 53.0]
Low immediacy robot	36.3	[33.5, 39.1]	46.5	[44.2, 48.8]
Human	41.5	[38.4, 44.5]	49.7	[47.0, 52.4]

6.2 Nonverbal Immediacy Results

Nonverbal immediacy scores were calculated from the questionnaires and produce a number which can be between 16 and 80. Immediacy scores and confidence intervals can be seen for each condition in Table 3. Whilst these scores might initially appear to be relatively low given the possibility of scores as high as 80, the scores do fall in the range expected. Due to the exclusion of certain aspects of the immediacy inventory in the robot conditions in terms of moving towards and touching observers, as well as producing facial expressions, it is unlikely that the score would raise above 56. It is however possible to be perceived differently and score more highly (for example the robot could have been perceived to have produced a smile, even though the mouth cannot move).

A two-tailed *t* test on the adult data reveals a significant difference between the nonverbal immediacy score for the high immediacy robot (M = 50.2, 95 % CI [47.0,53.5]) and the low immediacy robot (M = 36.3, 95 % CI [33.5,39.1]); $t(16) = 7.460$, $p < .001$. The same test on the child data also reveals a significant difference between the nonverbal immediacy score for the high immediacy robot (M = 50.8, 95 % CI [48.6,53.0]) and the low immediacy robot (M = 46.5, 95 % CI [44.2,48.8]); $t(53) = 2.793$, $p = .007$ (Fig. 4). These results confirm hypothesis H1, that robot behaviour designed to be more or less immediate will be perceived as such when measured using the nonverbal immediacy scale. This provides a useful check that the behaviour of the robot has been interpreted as intended by both children and adults.

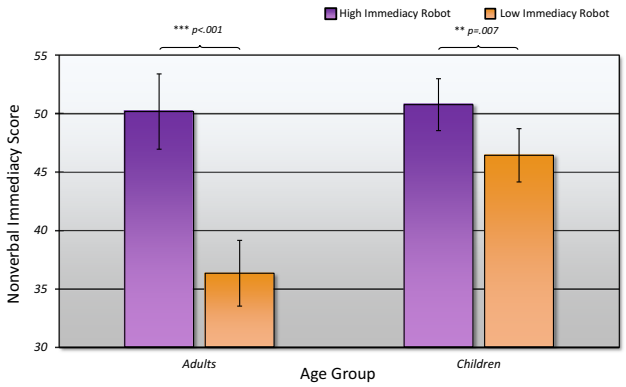


Fig. 4 Robot nonverbal immediacy scores as rated by children and adults, relating to hypothesis H1. Significance is indicated by $*p < .05$, $**p < .01$, and $***p < .001$. *Error bars* show the 95 % Confidence Interval

Support can be seen for hypothesis H2, that children and adults will perceive nonverbal immediacy in the same manner for both robots and humans (Table 3). The results show that both children and adults score the high immediacy robot very similarly, with almost identical means. The relative ranking of immediacy between conditions is also the same, with the high immediacy robot being perceived as most immediate, then the human, followed by the low immediacy robot condition.

However, there are also some differences as the child scores are more tightly bunched together; this could reflect their different (yet consistent) interpretation of negatively formulated questions [102], or more limited language understanding impeding the data quality [103]. A two-way ANOVA was conducted to examine the effect of age group (child/adult) and condition (high/low robot, human) on the immediacy rating. A significant interaction effect was found between these two factors: $F(2,108) = 5.29$, $p = .006$. Significant main effects were found for condition ($F(2,108) = 16.96$, $p < .001$) and age ($F(1,108) = 26.51$, $p < .001$). However, due to the interaction effect, exploration of simple main effects splitting the conditions is also required to correctly interpret the results. Significant simple main effects are found for condition within each level of age group (child/adult): adults—Wilks' Lambda = .796, $F(4,214) = 6.46$, $p < .001$; children—Wilks' Lambda = .798, $F(4,214) = 6.38$, $p < .001$. Significant simple main effects are also found for age group (child/adult) within each condition: low immediacy robot—Wilks' Lambda = .664, $F(2,107) = 27.11$, $p < .001$; high immediacy robot—Wilks' Lambda = .862, $F(2,107) = 8.54$, $p < .001$; human—Wilks' Lambda = .811, $F(2,107) = 12.49$, $p < .001$.

These findings suggest that some differences are present in the way that children perceive (or at least report) the immediacy of the characters when compared to adults. This is not surprising given the tighter bunching of child nonverbal immediacy scores. Nevertheless, there is a strong positive correlation between the child scores and the adult scores, $r(1) = 0.91$, although this is not significant ($p = .272$) due to the low number of comparisons (3 conditions). Overall, due to the strong positive correlation and the same ranking of the conditions, it would seem that children perceive nonverbal immediacy in a similar manner as adults, but there are clearly some differences at least in terms of reporting. We

would argue that there is a strong enough link to deem nonverbal immediacy an appropriate measure to use with children (and to tie the findings here to the adult human immediacy literature), but this is an area that would benefit from further research.

Cronbach's alpha values were calculated for the nonverbal immediacy questionnaire for adults and children, splitting the human condition and the robot conditions. All alpha values are based on the 16 item scale. The reliability rating for the adults with the robot is high ($\alpha = .79$), whereas in the human condition it is quite a bit lower ($\alpha = .45$). This difference may be an effect of embodiment, and will be explored further in the discussion Sect. 7.4. Reliability scores for children are relatively low in both cases (human $\alpha = .55$; robot $\alpha = .30$). In spite of the variation in child responses, the questionnaire was sensitive enough to detect differences as shown in this section. The implications of this are also discussed in Sect. 7.4.

6.3 Recall Results

Recall results are based on the 10 recall questions presented to all participants; scores are given as the correct proportion of answers, i.e. 8 correct answers = 0.8. Recall scores and confidence intervals can be seen for each condition in Table 4 and are represented graphically in Fig. 5.

To explore hypothesis H3, a two-tailed t test was conducted on the adult data to compare recall between observing the high and low immediacy robot conditions. No significant differences at the $p < .05$ level were found; $t(16) = -0.577$, $p = .572$. However, significant differences are found for the child data. A two-tailed independent samples t test reveals that recall is higher in the high immediacy robot condition ($M = 0.58$, 95 % CI [0.52,0.64]) than in the low immediacy robot condition ($M = 0.49$, 95 % CI [0.46,0.53]); $t(53) = 2.006$, $p = .011$.

These results provide partial support for hypothesis H3: recall will be greater when the character reading the story is more nonverbally immediate. It can be seen that this holds true for the children, where recall is greater in the high immediacy robot condition than in the low immediacy robot condition, in accordance with this condition being perceived as more immediate. However, there are no significant differences in recall between the conditions for adults. This is likely due to a ceiling effect with adults because the recall questions were designed so that they were suitable for children. This may have made them too easy for adults overall, leaving limited space to show differences between conditions. If the questions were more difficult and exclusively targeted towards adults then it is possible that differences would be found. The partial support for H3 and replication of findings from previous studies of nonverbal immediacy—using robots—provides a proof-of-concept for the approach proposed in this paper.

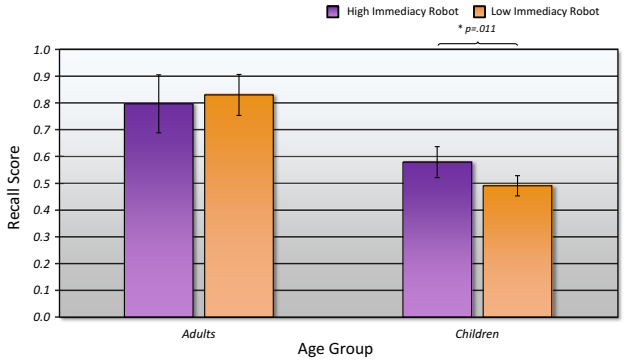


Fig. 5 Recall scores for high and low nonverbal immediacy robot conditions relating to hypothesis H3. Significance is indicated by $*p < .05$, $**p < .01$, and $***p < .001$. *Error bars* show the 95 % Confidence Interval

No support is found for hypothesis H4: that higher individual perception of nonverbal immediacy will lead to greater recall for that individual. Correlations between nonverbal immediacy ratings and recall scores are not significant for children ($r(81) = -0.047$; $p = .673$) or adults ($r(29) = -0.188$; $p = .311$). Indeed the correlations themselves are in the opposite direction (although only with a small magnitude) to that which was expected. This would suggest that in this study, the rating of immediacy at the individual level has less of a bearing on recall than the average as judged by the group, but there is not enough evidence here to explain why this occurred.

7 Discussion

This paper started from the established research field of *nonverbal immediacy* which links behaviour to learning gains in a measurable and comparable manner (Sect. 2). This was broken down into its component social cues to explore their

Table 4 Mean recall scores by condition

Condition	Adult M	95 % CI	Child M	95 % CI
High immediacy robot	0.80	[0.69, 0.91]	0.58	[0.52, 0.64]
Low immediacy robot	0.83	[0.76, 0.91]	0.49	[0.46, 0.53]
Human	0.79	[0.73, 0.84]	0.63	[0.56, 0.70]

effect on learning individually. The evaluation in this paper applied a series of guidelines that were devised based on nonverbal immediacy cues and informed by HHI and HRI literature. It was found that both children and adults perceive the immediacy of a robot designed to have low and high nonverbal immediacy behaviours as intended, which confirms and extends prior work in HRI [23]. Additionally, both children and adults ranked the nonverbal immediacy of robots and humans in the same order, although children's raw scores were more tightly grouped. This gives rise to the possibility that much of the nonverbal immediacy literature, which has mostly been conducted with adults, would also apply to children.

Recall of a short story improved significantly for children when the robot reading the story was more immediate in behaviour, which does indeed confirm the hypothesis derived from nonverbal immediacy literature, based on human–human studies showing the same effect [37,38]. No significant difference in recall was observed in the adult data, but this may be due to the relative lack of difficulty of the recall test, which had been designed specifically for children.

The following subsections will discuss the findings here in the wider context of research conducted in HRI and HHI. First the impact of individual characteristics will be discussed in relation to hypothesis H4, which was not supported. Secondly, the possible impact of novelty on the perception of behaviour and recall will be explored. Thirdly, potential shortcomings of nonverbal immediacy as a measure for characterising interactions are raised. Finally, we share the lessons learnt from this study in applying nonverbal immediacy measures to HRI and consider the influence of the study design on the findings.

7.1 Students as Individuals

Out of necessity, most experiments observe the learning of large samples of students, meaning that the effect is seen on average, but does not necessarily apply to all students. All children are individuals, with their own characteristics, preferences for subjects and learning styles. It may be that there are some educational scenarios, topics, or children, with which technology is more suited to assisting [104]. Some children may be impacted to a degree related to their personality (and their 'need to belong') [105], or their learning style [106], which can affect their sensitivity to social cues.

All studies here have been considering typically developing children/students, so many of the outcomes may not apply to individuals with, for example, attention-deficit hyperactive disorder (ADHD) or autism spectrum disorder (ASD) who might have difficulties in interpreting some social cues [107–109].

Gender could also have an impact on learning and the use of social cues. It has been found in both virtual environments [110–112] and physical environments [113] that males do not utilise gaze cues in the same way as females; or if they do, it does not manifest in behaviour change or learning. The gender of the teacher, at least in virtual environments, does not however seem to impact on the learning which takes place [114].

In the evaluation presented in Sect. 6, support was not found for hypothesis H4, which sought to link individual perceptions of the robot behaviour (as measured through nonverbal immediacy) to recall scores. It is suggested that this may be because the nonverbal immediacy scale does not cater for the many other variables between individuals that may influence their learning. However, this does not reduce the utility of nonverbal immediacy as a characterisation of robot social behaviour, with differences in robot behaviours clearly demonstrated as part of hypothesis H1. Instead, we highlight here the need to further develop means of including perceptions of robot behaviour into broader models of learner characteristics.

7.2 The Novelty Aspect

It is necessary to acknowledge that the use of social cues is only partially responsible for positive learning outcomes. The approach, content and assessment of teaching contributes significantly to the learning process [115], as does the knowledge of the teacher [116] and their beliefs towards learning [117]. Of course, the students play an equal part in learning too, with aspects such as their emotion playing a role in the process [118]. Teachers and students often have long-standing relationships; these relationships allow for familiarisation with teaching and learning styles, which is beneficial for learning: when teacher turnover increases, attainment scores have been shown to drop, evidencing the importance of consistent relationships [119]. This highlights the need for long-term interaction if using social robots to assist in education, alongside thorough development of learning materials.

The majority of the studies considered as part of the analysis conducted here only look at single interactions, rather than interactions over time. There is evidence for changing preferences (and thus possibly changes in subsequent learning outcomes) over time, as seen in [120]. Of course, a relative lack of long-term data in HRI is understandable because of the immense challenge in enforcing methodological rigour over extended periods of time and the ethical implications of using atypical conditions (such as the low immediacy robot condition from the evaluation in this paper) in real-world learning.

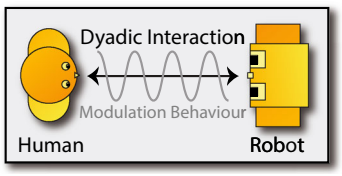


Fig. 6 Representation of the role of social cues in dyadic HRI. Social cues are used as modulation behaviour within the interaction

7.3 Nonverbal Immediacy and Interaction

Due to the potentially great benefits of using robots as tutors in one-on-one interactions [121,122], and the possibility of personalisation in such contexts, this seems to be an apt means of applying robots in education. Whilst nonverbal immediacy addresses how competent a speaker is at communicating towards others, i.e. how well a teacher can convey information to students, in one-to-one tutoring it is important to be competent at two-way communication as well. As such, it may be that the approach taken in this paper would need adapting for one-to-one tutoring, incorporating more principles from dyadic interaction work.

Social behaviour plays a key role in dyadic interaction and on the outcome of communication within a dyad. The role of communication, or the social interaction within the dyad, in such a scenario is posited to be "the mutual modification of two ongoing streams of behaviour of two persons" [123]. The behaviour of one party affects the behaviour of the other. In this view, social cues are used as part of the modulating behaviour in this process (Fig. 6) and can therefore be utilised in many processes influencing education.

The joint modification of behaviour within the dyad gives rise to the need for regulation and alignment of behaviour in order to simultaneously transmit and receive information [124]. All parties engaging in a social interaction must continually adapt the social cues they are using in order to effectively construct the interaction [125]; for example, verbal turn-taking must be regulated through the use of various social cues [123]. Such regulation is important in learning interactions, indicating when it is appropriate for learners to ask questions, and when it is time for them to receive information; learning is more challenging without social cues or conventions to manage this turn-taking [126]. This simple coordination in interaction is vital and has been shown to influence cognition from infancy [124]. Even in unstructured interactions with robots, children appear to actively seek such turn-taking in interactions [127].

These kinds of interaction phenomena are not catered for in nonverbal immediacy measures. The evaluation in this paper saw positive results, but the interaction between the robot and the humans was largely in one direction (the robot instructing the humans); the robot was not responsive to human social cues or behaviour. This is an area which needs further exploration in HRI: the question is when the interaction becomes more interactional than those presentational behaviours considered in the present study, do immediacy principles hold, or are additional behaviours (such as turn-taking policies) required? We propose that in the absence of further evidence in such contexts, the application of the non-verbal immediacy metric provides a suitable basis for initial investigation.

7.4 Using Nonverbal Immediacy in HRI

Whilst the evaluation in this paper had positive results and confirmed (or partially confirmed) three of the four hypotheses, it should be made clear that there are limitations imposed by the study design which could inhibit how well these findings translate to other scenarios. The human condition was shown through a video, whereas the robots were physically present. This means that a comparison between the recall and nonverbal immediacy scores from the human and the robot conditions could be influenced by embodiment, or social facilitation effects [128]. It should be noted that in this study, we do not directly compare between these conditions: comparisons are made within robot conditions, or from children and adults, but not between the human and robot conditions.

The reliability metrics across the conditions demonstrate the effectiveness of the nonverbal immediacy characterisation of social behaviour. Generally, the adult raters have high reliability levels, which reflects the behaviour seen in the literature. That this applies to ratings of robot behaviours indicates the applicability of the metric. Whereas the alpha statistic is lower for children, there are two points of note. Firstly, there remains a reasonable consistency for the ratings of the human condition—this extends the literature by showing the ability of children (in addition to adults) to use the nonverbal immediacy metric. Secondly, for both children and adults, there was agreement in the ordering of relative immediacy levels between the conditions—this indicates that the non-verbal immediacy scale is sensitive enough for the present study, for both adults and children.

A number of caveats apply however that require further investigation. A high reliability score is found for the adults who saw a robot condition, but this is not so high for those who saw the human condition. This may be due to relatively low subject numbers when considering only the human condition (13 subjects), where inconsistency from one or two individual subjects could have a large impact on the alpha value. The reliability for the human is higher for children than for adults, suggesting the difference in subject numbers could be a factor. Alternatively, it could be a result of embodiment: the robot conditions were seen in person, whereas the human was shown on screen, which may have influenced the reporting of social behaviour on the questionnaire.

The Cronbach's alpha statistic for the children who saw a robot condition is considerably lower than that of the adults. This is not so surprising, given the complications highlighted in the literature of using questionnaires with children [103]. However, it may also be a product of limitations in robot social behaviour. Cronbach's alpha measures the internal consistency of questionnaire items. Whilst some inconsistency is likely due in part to child interpretations of negatively worded items [102], there are some items within the questionnaire that the robot behaviour itself is probably not consistent in. For example, the questions related to smiling and frowning are opposites of each other in terms of calculating a value for the scale, but could both be answered as 'never' performed, as the robot does not have moveable facial features. Such a response would provide maximum inconsistency between these items. This would not necessarily reflect the reliability of the questionnaire, but a limitation in the ability of the robot to implement all of the questionnaire items. The same argument could be made for the items concerning touch—it could be considered that the robot never touches the observer, whilst also not 'avoiding' touch, as the question is worded. Inclusion of these two behavioural elements (that were not possible in the evaluation here) in subsequent work exploring the use of nonverbal immediacy for characterising robot social behaviour would likely yield higher reliability scores.

The interaction was also over a very short period of time (approximately 4–5 min) and the measurement of learning was through recall. Although recall is a fundamental element of learning, it is very different from understanding or applying knowledge, or from the higher dimensions of learning as defined in the revised version of Bloom's taxonomy [29]. Early results suggest that nonverbal immediacy can also be applied in slightly longer interactions, and in dyadic contexts, with learning positively improved as nonverbal immediacy increases [18]. However, longer scale studies with a variety of robots and learning materials would certainly add more weight to the evidence of how well nonverbal immediacy can be applied to HRI.

8 Conclusion

This paper introduced a variety of literature from the well-established area of research studying nonverbal immediacy. Nonverbal immediacy can be used to characterise social behaviour through observer-reports on the use of social cues, such as gaze and gesture. We explored HHI and HRI literature relating to these cues and brought the findings together into a set of guidelines for robot social behaviour. These guidelines were implemented in an evaluation that compared an intended high nonverbal immediacy and a low nonverbal immediacy robot. A human condition was also included to link the work here to existing nonverbal immediacy literature and provide validation for the use of nonverbal immediacy with children. Several hypotheses derived from the nonverbal immediacy literature were confirmed. Both children and adults judge the immediacy of humans and robots in a similar manner. The children's responses were more varied than the adults, but it was still possible to identify a significant difference in their perception of the social behaviour between the two robot conditions. Children also recalled more of the story when the robot used more nonverbal immediacy behaviours, which demonstrates an effect predicted by the literature. While there are some limitations in the measure, it is proposed that nonverbal immediacy could be used as an effective means of characterising robot social behaviour for human–robot interaction, for both adult and child subjects.

Acknowledgements This research was partially funded by the EU FP7 DREAM project (FP7-ICT-611391) and the School of Computing and Maths, Plymouth University, UK. Thanks goes to CAEN Community Primary School, Braunton, UK. for taking part in the evaluation.

Appendix 1: Short Story Script

The following is the short story script as used in all evaluation conditions. The story is largely based on one from the following website: http://freestoriesforkids.com/children/stories-and-tales/robot-virus (produced here with permission from the author).

Hello, I'm Charlie. Today I'm going to tell you one of my favourite robot stories. It is about a boy, his name is Ricky, and his robot helper, Johnny. Ricky lived in a lovely futuristic house, which had everything you could ever want. Though he didn't help much around the house, Ricky was still as pleased as punch when his parents bought him the latest model of helper robot. As soon as it arrived, off it went; cooking, cleaning, ironing, and—most importantly—gathering up old clothes from Ricky's bedroom floor, which Ricky didn't like having to walk on.

On that first day, when Ricky went to sleep, he had left his bedroom in a truly disastrous state. When he woke up the next morning, everything was perfectly clean and tidy. In fact, it was actually too clean. Ricky could not find his favourite blue skateboard. However much he searched, it did not reappear, and the same was starting to happen with other things. Ricky looked with suspicion at the gleaming helper robot. He hatched a plan to spy on the robot, and began following it around the house.

Finally he caught it red-handed. It was picking up a toy to hide it. Off he went, running to his parents, to tell them that the helper was broken and badly programmed. Ricky asked them to have it changed. But his parents said absolutely not; it was impossible, they were delighted with the new helper, and that it was the best cleaner they had ever met. So Ricky needed to

get some kind of proof; maybe take some hidden photos. He kept nagging his parents for three whole weeks about how much good stuff the robot was hiding. Ricky argued that this was not worth the clean house because toys are more important.

One day the robot was whirring past, and heard the boy's complaints. The robot returned with five of his toys, and some clothes for him."Here sire, I did not know it was bothering you", said the helper, with its metallic voice. "How could it not you thief?! You've been nicking my stuff for weeks", the boy answered, furiously. The robot replied, "the objects were left on the floor. I therefore calculated that you did not like them. I am programmed to collect all that is not wanted, and at night I send it to places other humans can use it. I am a maximum efficiency machine. Did you not know?".

Ricky started feeling ashamed. He had spent all his life treating things as though they were useless. He looked after nothing. Yet it was true that many other people would be delighted to treat those things with all the care in the world. And he understood that the robot was neither broken nor badly programmed, rather, it had been programmed extremely well! Since then, Ricky decided to become a Maximum Efficiency Boy, and he put real care into how he treated his things. He kept them tidy, and made sure that he didn't have more than was necessary. And, often, he would buy things, and take them along with his good friend, the robot, to help out those other people who needed them.

The end... I hope you enjoyed the story. Goodbye!

Appendix 2: Robot Nonverbal Immediacy Questionnaire (RNIQ)

The following is the questionnaire used by participants in the evaluation to rate the nonverbal immediacy of the robot, as based on the short-form nonverbial immediacy scale-observer report. The directions are provided verbally by the experimenter, so the top of the survey simply asks to 'please put a circle around your choice for each question'. Options are provided in equally sized boxes below each question. The options are: *1 = Never; 2 = Rarely; 3 = Sometimes; 4 = Often; 5 = Very Often*. The questions are as follows:

1. The robot uses its hands and arms to gesture while talking to you
2. The robot uses a dull voice while talking to you
3. The robot looks at you while talking to you
4. The robot frowns while talking to you
5. The robot has a very tense body position while talking to you
6. The robot moves away from you while talking to you
7. The robot varies how it speaks while talking to you
8. The robot touches you on the shoulder or arm while talking to you
9. The robot smiles while talking to you
10. The robot looks away from you while talking to you
11. The robot has a relaxed body position while talking to you
12. The robot stays still while talking to you
13. The robot avoids touching you while talking to you
14. The robot moves closer to you while talking to you
15. The robot looks keen while talking to you
16. The robot is bored while talking to you

Scoring

Step 1 Add the scores from the following items:
1, 3, 7, 8, 9, 11, 14, and 15.
Step 2 Add the scores from the following items:
2, 4, 5, 6, 10, 12, 13, and 16.
Total Score = 48 plus Step 1 minus Step 2.

This questionnaire can also be downloaded online.[6] The online version has been modified from the version shown here as children commonly did not understand the word 'varies' in question 7, so this now reads 'changes'.

Appendix 3: Recall Quesionnaire

The following questions are those used in the recall questionnaire; in brackets after each question are the possible answers.

1. What is the name of the boy in the story? {Ricky, Mickey, Harry, Jeff}
2. What is the name of the robot in the story? {Rupert, John, Johnny, George}
3. What was the most important thing for the robot to pick up from the floor of the boy's bedroom? {clothes, food, toys, t-shirts}
4. What did the boy think about doing to get proof of the robot taking his things? {taking photos, shouting at it, taking video, telling his parents}
5. What toy couldn't the boy find the first day after the robot had tidied? {orange skateboard, games console, blue skateboard, blue doll}
6. How many toys did the robot give back to the boy after he complained? {eight (8), five (5), three (3), six (6)}
7. How long did the boy complain to his parents for? {three (3) weeks, eight (8) days, three (3) days, four (4) weeks}
8. What type of boy did he decide to be at the end of the story? {maximum efficiency, tidy, minimum efficiency, messy}

[6] http://www.tech.plym.ac.uk/SoCCE/CRNS/staff/JKennedy/Robot_Nonverbal_Immediacy_Questionnaire.

9. What type of robot is the one in the story? {angry, purple, helper, flying}
10. What is the robot in the story especially good at? {ironing, swimming, jumping, cleaning}
11. What was the moral of the story? free text answer

References

1. Gordon G, Breazeal C, Engel S (2015) Can children catch curiosity from a social robot? In: Proceedings of the 10th ACM/IEEE international conference on human-robot interaction, ACM
2. Kennedy J, Baxter P, Belpaeme T (2015c) The robot who tried too hard: social behaviour of a robot tutor can negatively affect child learning. In: Proceedings of the 10th ACM/IEEE international conference on human-robot interaction, ACM, pp 67–74. doi:10.1145/2696454.2696457
3. Short E, Swift-Spong K, Greczek J, Ramachandran A, Litoiu A, Grigore EC, Feil-Seifer D, Shuster S, Lee JJ, Huang S, Levonisova S, Litz S, Li J, Ragusa G, Spruijt-Metz D, Matarić M, Scassellati B (2014) How to train your DragonBot: Socially assistive robots for teaching children about nutrition through play. In: Proceedings of the 23rd IEEE international symposium on robot and human interactive communication, IEEE, RO-MAN, 2014, pp 924–929
4. Alemi M, Meghdari A, Ghazisaedy M (2014) Employing humanoid robots for teaching english language in Iranian junior high-schools. Int J Hum Robot. doi:10.1142/S0219843614500224
5. Leite I, McCoy M, Lohani M, Ullman D, Salomons N, Stokes C, Rivers S, Scassellati B (2015) Emotional storytelling in the classroom: individual versus group interaction between children and robots. In: Proceedings of the 10th annual ACM/IEEE international conference on human-robot interaction, ACM, pp 75–82
6. Kyriakides L, Creemers BP, Antoniou P (2009) Teacher behaviour and student outcomes: suggestions for research on teacher training and professional development. Teach Teach Educ 25(1):12–23
7. Saerbeck M, Schut T, Bartneck C, Janse MD (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, ACM, New York, NY, USA, CHI'10, pp 1613–1622. doi:10.1145/1753326.1753567
8. Blanson Henkemans OA, Bierman BP, Janssen J, Neerincx MA, Looije R, van der Bosch H, van der Giessen JA (2013) Using a robot to personalise health education for children with diabetes type 1: a pilot study. Patient Educ Couns 92(2):174–181
9. Leyzberg D, Spaulding S, Scassellati B (2014) Personalizing robot tutors to individual learning differences. In: Proceedings of the 9th ACM/IEEE international conference on human-robot interaction
10. Janssen J, van der Wal C, Neerincx M, Looije R (2011) Motivating children to learn arithmetic with an adaptive robot game. Soc Robot 153–162
11. Bandura A, McClelland DC (1977) Social learning theory. Prentice-Hall, Englewood Cliffs, NJ
12. Vygotsky LS (1980) Mind in society: the development of higher psychological processes. Harvard University Press, Cambridge
13. Wu R, Kirkham NZ (2010) No two cues are alike: depth of learning during infancy is dependent on what orients attention. J Exp Child Psychol 107(2):118–136
14. Roth WM, Lawless DV (2002) When up is down and down is up: body orientation, proximity, and gestures as resources. Lang Soc 31(01):1–28
15. Strong M, Gargani J, Hacifazlioğlu Ö (2011) Do we know a successful teacher when we see one? experiments in the identification of effective teachers. J Teach Educ 62(4):367–382
16. Kreijns K, Kirschner PA, Jochems W (2003) Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. Comput Hum Behav 19(3):335–353
17. Bloom B, Engelhart M, Furst E, Hill W, Krathwohl D (1956) Taxonomy of educational objectives: the classification of educational goals. Handbook I: cognitive domain. Donald McKay, New York
18. Kennedy J, Baxter P, Senft E, Belpaeme T (2015d) Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions. In: International conference on social robotics
19. Castellano G, Paiva A, Kappas A, Aylett R, Hastie H, Barendregt W, Nabais F, Bull S (2013) Towards empathic virtual and robotic tutors. Artificial Intelligence in Education. Springer, New York, pp 733–736. doi:10.1007/978-3-642-39112-5_100
20. Sharma M, Hildebrandt D, Newman G, Young JE, Eskicioglu R (2013) Communicating affect via flight path: exploring use of the laban effort system for designing affective locomotion paths. In: Proceedings of the 8th ACM/IEEE international conference on human-robot interaction, HRI '13, pp 293–300
21. Andrist S, Spannan E, 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, IEEE Press, pp 341–348
22. Cramer HS, Kemper NA, Amin A, Evers V (2009) The effects of robot touch and proactive behaviour on perceptions of human-robot interactions. In: Proceedings of the 4th ACM/IEEE international conference on human robot interaction, ACM, pp 275–276
23. Szafir D, Mutlu B (2012) Pay Attention!: designing adaptive agents that monitor and improve user engagement. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, New York, NY, USA, CHI'12, pp 11–20. doi:10.1145/2207676.2207679
24. Zaki J (2013) Cue integration a common framework for social cognition and physical perception. Perspectives on Psychological Science 8(3):296–312
25. Mehrabian A (1968) Some referents and measures of nonverbal behavior. behav res methods instrum 1(6):203–207
26. Richmond VP, McCroskey JC, Johnson AD (2003) Development of the nonverbal immediacy scale (NIS): measures of self- and other-perceived nonverbal immediacy. Commun Q 51(4):504–517
27. Witt PL, Wheeless LR, Allen M (2004) A meta-analytical review of the relationship between teacher immediacy and student learning. Commun Monogr 71(2):184–207
28. Krathwohl D, Bloom B, Masia B (1964) Taxonomy of educational objectives: The classification of educational goals. Handbook II: the affective domain. Donald McKay, New York
29. Krathwohl DR (2002) A revision of bloom's taxonomy: an overview. Theory Pract 41(4):212–218
30. Gorham J (1988) The relationship between verbal teacher immediacy behaviors and student learning. Commun Educ 37(1):40–53
31. Chesebro JL, McCroskey JC (2000) The relationship between students' reports of learning and their actual recall of lecture material: a validity test. Commun Educ 49(3):297–301
32. McCroskey JC, Sallinen A, Fayer JM, Richmond VP, Barraclough RA (1996) Nonverbal immediacy and cognitive learning: a cross-cultural investigation. Commun Educ 45(3):200–211
33. Kennedy J, Baxter P, Belpaeme T (2015a) Can less be more? The impact of robot social behaviour on human learning. In: Proceedings of the 4th international symposium on new frontiers in HRI at AISB 2015
34. Christensen LJ, Menzel KE (1998) The linear relationship between student reports of teacher immediacy behaviors and perceptions of state motivation, and of cognitive, affective, and

behavioral learning. Commun Educ 47(1):82–90. doi:10.1080/03634529809379112
35. Christophel DM (1990) The relationships among teacher immediacy behaviors, student motivation, and learning. Commun Educ 39(4):323–340
36. Chesebro JL (2003) Effects of teacher clarity and nonverbal immediacy on student learning, receiver apprehension, and affect. Commun Educ 52(2):135–147
37. Goodboy AK, Weber K, Bolkan S (2009) The effects of nonverbal and verbal immediacy on recall and multiple student learning indicators. J Classr Interact 44(1):4–12
38. Witt PL, Wheeless LR (2001) An experimental study of teachers' verbal and nonverbal immediacy and students' affective and cognitive learning. Commun Educ 50(4):327–342. doi:10.1080/03634520109379259
39. Chesebro JL, McCroskey JC (1998) The relationship of teacher clarity and teacher immediacy with students experiences of state receiver apprehension. Commun Q 46(4):446–456
40. Chidambaram V, Chiang YH, Mutlu B (2012) Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, ACM, pp 293–300
41. Jeong S, Gu J, Shin DH (2015) I am interested in what you are saying: role of nonverbal immediacy cues in listening. In: Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction extended abstracts, ACM, pp 129–130
42. Comstock J, Rowell E, Bowers JW (1995) Food for thought: teacher nonverbal immediacy, student learning, and curvilinearity. Commun Educ 44(3):251–266
43. Witt PL, Schrodt P (2006) The influence of instructional technology use and teacher immediacy on student affect for teacher and course. Commun Rep 19(1):1–15
44. Kelly SD, Manning SM, Rodak S (2008) Gesture gives a hand to language and learning: perspectives from cognitive neuroscience, developmental psychology and education. Lang Linguist Compass 2(4):569–588
45. Macedonia M, von Kriegstein K (2012) Gestures enhance foreign language learning. Biolinguistics 6(3–4):393–416
46. Goldin-Meadow S, Wagner SM (2005) How our hands help us learn. Trends Cogn Sci 9(5):234–241
47. Goldin-Meadow S, Kim S, Singer M (1999) What the teacher's hands tell the student's mind about math. J Educ Psychol 91(4):720–730. doi:10.1037/0022-0663.91.4.720
48. Goodwyn SW, Acredolo LP (1998) Encouraging symbolic gestures: A new perspective on the relationship between gesture and speech. New Dir Child Adolesc Dev 79:61–73
49. Goldin-Meadow S, Nusbaum H, Kelly SD, Wagner S (2001) Explaining math: gesturing lightens the load. Psychol Sci 12(6):516–522
50. Cook SW, Mitchell Z, Goldin-Meadow S (2008) Gesturing makes learning last. Cognition 106(2):1047–1058
51. Goldin-Meadow S, Wein D, Chang C (1992) Assessing knowledge through gesture: using children's hands to read their minds. Cogn Instr 9(3):201–219
52. Valenzeno L, Alibali MW, Klatzky R (2003) Teachers gestures facilitate students learning: a lesson in symmetry. Contemp Educ Psychol 28(2):187–204
53. Roth WM (2001) Gestures: their role in teaching and learning. Rev Educ Res 71(3):365–392
54. Rumme P, Saito H, Ito H, Oi M, Lepe A (2008) Gestures as effective teaching tools: are students getting the point? In: Japanese Cognitive Science Society Meeting 2008
55. Wang X, Williams MA, Gardenfors P, Vitale J, Abidi S, Johnston B, Kuipers B, Huang A (2014) Directing human attention with pointing. In: 23rd IEEE international symposium on IEEE robot and human interactive communication, 2014 RO-MAN, pp 174–179
56. Gullberg M, Holmqvist K (2002) Visual attention towards gestures in face-to-face interaction vs. on screen. In: Wachsmuth I, Sowa T (eds) Gesture and sign language in human-computer interaction. Springer, Berlin, pp 206–214
57. Wu R, Gopnik A, Richardson DC, Kirkham NZ (2010) Social cues support learning about objects from statistics in infancy. In: Proceedings of the 32nd annual conference of the cognitive science society, pp 1228–1233
58. Yu C, Ballard DH (2007) A unified model of early word learning: integrating statistical and social cues. Neurocomputing 70(13):2149–2165
59. Houston-Price C, Plunkett K, Duffy H (2006) The use of social and salience cues in early word learning. J Exp Child Psychol 95(1):27–55
60. Powell KL, Roberts G, Nettle D (2012) Eye images increase charitable donations: evidence from an opportunistic field experiment in a supermarket. Ethology 118(11):1096–1101
61. Boucher JD, Ventre-Dominey J, Dominey PF, Fagel S, Bailly G (2010) Facilitative effects of communicative gaze and speech in human-robot cooperation. In: Proceedings of the 3rd international workshop on affective interaction in natural environments, ACM, New York, NY, USA, AFFINE '10, pp 71–74. doi:10.1145/1877826.1877845
62. Okumura Y, Kanakogi Y, Kanda T, Ishiguro H, Itakura S (2013) The power of human gaze on infant learning. Cognition 128(2):127–133
63. Admoni H, Bank C, Tan J, Toneva M, Scassellati B (2011) Robot gaze does not reflexively cue human attention. In: Processings of the 33rd annual conference of the cognitive science society (2011), pp 1983–1988
64. Sherwood JV (1987) Facilitative effects of gaze upon learning. Percept Mot Skills 64(3c):1275–1278
65. Otteson JP, Otteson CR (1979) Effect of teacher's gaze on children's story recall. Percept Mot Skills 50(1):35–42
66. Dalzel-Job O, Oberlander J, Smith TJ (2011) Don't look now: the relationship between mutual gaze, task performance and staring in second life. In: Proceedings of the 33rd annual conference of the cognitive science society, pp 832–837
67. Kennedy J, Baxter P, Belpaeme T (2015b) Comparing robot embodiments in a guided discovery learning interaction with children. Int J Social Robot 7(2):293–308. doi:10.1007/s12369-014-0277-4
68. Looije R, van der Zalm A, Neerincx MA, Beun RJ (2012) Help, I need some body the effect of embodiment on playful learning. In: The 21st IEEE international symposium on robot and human interactive communication, IEEE, RO-MAN 2012, pp 718–724. doi:10.1109/ROMAN.2012.6343836
69. Atkinson RK, Mayer RE, Merrill MM (2005) Fostering social agency in multimedia learning: examining the impact of an animated agents voice. Contemp Educ Psychol 30(1):117–139
70. Mayer RE, Sobko K, Mautone PD (2003) Social cues in multimedia learning: role of speaker's voice. J Educ Psychol 95(2):419–425
71. Mori M, MacDorman KF, Kageki N (2012) The uncanny valley [from the field]. IEEE Robot Autom Mag 19(2):98–100
72. Baylor A, Ryu J, Shen E (2003) The effects of pedagogical agent voice and animation on learning, motivation and perceived persona. In: World conference on educational multimedia, hypermedia and telecommunications, pp 452–458
73. Remland MS, Jones TS (1994) The influence of vocal intensity and touch on compliance gaining. J Soc Psychol 134(1):89–97
74. Simonds BK, Meyer KR, Quinlan MM, Hunt SK (2006) Effects of instructor speech rate on student affective learning, recall, and

perceptions of nonverbal immediacy, credibility, and clarity. Commun Res Rep 23(3):187–197. doi:10.1080/08824090600796401
75. Fulford CP (1992) Systematically designed text enhanced with compressed speech audio. In: Proceedings of selected research and development presentations at the convention of the association for educational communications and technology
76. Velez JJ, Cano J (2008) The relationship between teacher immediacy and student motivation. J Agric Educ 49(3):76–86
77. Becker-Asano C, Stahl P, Ragni M, Courgeon M, Martin JC, Nebel B (2013) An affective virtual agent providing embodied feedback in the paired associate task: system design and evaluation. In: Intelligent Virtual Agents, Springer, pp 406–415
78. Peters P (2007) Gaining compliance through non-verbal communication. Pepperdine Dispute Resolut Law J 7(1):87–112
79. Segrin C (1993) The effects of nonverbal behavior on outcomes of compliance gaining attempts. Commun Stud 44(3–4):169–187
80. Greene LR (1977) Effects of verbal evaluation feedback and interpersonal distance on behavioral compliance. J Couns Psychol 24(1):10
81. Hiroi Y, Ito A (2011) Influence of the size factor of a mobile robot moving toward a human on subjective acceptable distance. Mob Robots Curr Trends. doi:10.5772/26512
82. Kim Y, Mutlu B (2014) How social distance shapes humanrobot interaction. Int J Hum Comput Stud 72(12):783–795. doi:10.1016/j.ijhcs.2014.05.005
83. Walters ML, Dautenhahn K, Te Boekhorst R, Koay KL, Kaouri C, Woods S, Nehaniv C, Lee D, Werry I (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, 2005. ROMAN 2005, IEEE, pp 347–352
84. Kennedy J, Baxter P, Belpaeme T (2014) Children comply with a robot's indirect requests. In: Proceedings of the 9th ACM/IEEE international conference on human-robot interaction, pp 198–199. doi:10.1145/2559636.2559820
85. Aiello JR, Aiello TDC (1974) The development of personal space: proxemic behavior of children 6 through 16. Hum Ecol 2(3):177–189
86. Huettenrauch H, Severinson Eklundh K, Green A, Topp E (2006) Investigating spatial relationships in human-robot interaction. In: IEEE/RSJ international conference on intelligent robots and systems, pp 5052–5059. doi:10.1109/IROS.2006.282535
87. Takayama L, Pantofaru C (2009) Influences on proxemic behaviors in human-robot interaction. In: IEEE/RSJ international conference on intelligent robots and systems, pp 5495–5502. doi:10.1109/IROS.2009.5354145
88. Mumm J, 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, ACM, HRI '11, pp 331–338. doi:10.1145/1957656.1957786
89. Rae I, Takayama L, Mutlu B (2013) The influence of height in robot-mediated communication. In: Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction, IEEE Press, pp 1–8
90. Fisher JD, Rytting M, Heslin R (1976) Hands touching hands: affective and evaluative effects of an interpersonal touch. Sociometry 39(4):416–421
91. Fukuda H, Shiomi M, Nakagawa K, Ueda K (2012) 'Midas touch' in human-robot interaction: evidence from event-related potentials during the ultimatum game. In: Proceedings of the 7th ACM/IEEE international conference on human-robot interaction, ACM, pp 131–132
92. Gurung RA, Vespia K (2007) Looking good, teaching well? linking liking, looks, and learning. Teach Psychol 34(1):5–10
93. Guéguen N (2002) Touch, awareness of touch, and compliance with a request. Perceptual and motor skills 95(2):355–360
94. Salter T, Dautenhahn K, te Boekhorst R (2006) Learning about natural human-robot interaction styles. Robot Auton Syst 54(2):127–134
95. Byrd CE, McNeil N, D'Mello S, Cook SW (2014) Gesturing may not always make learning last. In: Proceedings of the 36th annual conference of the cognitive science society, pp 1982–1987
96. Langton SR (2000) The mutual influence of gaze and head orientation in the analysis of social attention direction. Q J Exp Psychol A 53(3):825–845
97. Langton SR, Bruce V (2000) You must see the point: automatic processing of cues to the direction of social attention. J Exp Psychol Hum Percept Perform 26(2):747
98. Lohan KS, Rohlfing K, Saunders J, Nehaniv C, Wrede B (2012) Contingency scaffolds language learning. In: IEEE international conference on development and learning and epigenetic robotics, ICDL, pp 1–6
99. Anderson LW (1975) Student involvement in learning and school achievement. Calif J Educ Res 26(2):53–62
100. Richmond VP, McCroskey JC (1998) Nonverbal communication in interpersonal relationships, 3rd edn. Allyn and Bacon, Boston
101. Hulme C, Tordoff V (1989) Working memory development: the effects of speech rate, word length, and acoustic similarity on serial recall. J Exp Child Psychol 47(1):72–87. doi:10.1016/0022-0965(89)90063-5
102. Borgers N, Sikkel D, Hox J (2004) Response effects in surveys on children and adolescents: the effect of number of response options, negative wording, and neutral mid-point. Qual Quant 38(1):17–33
103. Borgers N, De Leeuw E, Hox J (2000) Children as respondents in survey research: cognitive development and response quality 1. Bulletin de methodologie Sociologique 66(1):60–75
104. Dede C (2009) Immersive interfaces for engagement and learning. Science 323(5910):66–69
105. Pickett CL, Gardner WL, Knowles M (2004) Getting a cue: the need to belong and enhanced sensitivity to social cues. Pers Soc Psychol Bull 30(9):1095–1107
106. Witkin HA, Moore CA, Goodenough DR, Cox PW (1977) Field-dependent and field-independent cognitive styles and their educational implications. Rev Educ Res 1–64
107. Bauminger N (2002) The facilitation of social-emotional understanding and social interaction in high-functioning children with autism: intervention outcomes. J Autism Dev Disord 32(4):283–298
108. Hall CW, Peterson AD, Webster RE, Bolen LM, Brown MB (1999) Perception of nonverbal social cues by regular education, ADHD, and ADHD/LD students. Psychol Sch 36(6):505–514
109. Jellema T, Lorteije J, van Rijn S, van t'Wout M, de Haan E, van Engeland H, Kemner C (2009) Involuntary interpretation of social cues is compromised in autism spectrum disorders. Autism Res 2(4):192–204
110. Bailenson J, Blascovich J, Beall A, Loomis J (2001) Equilibrium theory revisited: mutual gaze and personal space in virtual environments. Presence 10(6):583–598
111. Bailenson JN, Blascovich J, Beall AC, Loomis JM (2003) Interpersonal distance in immersive virtual environments. Pers Soc Psychol Bull 29(7):819–833
112. Bailenson JN, Beall AC, Loomis J, Blascovich J, Turk M (2005) Transformed social interaction, augmented gaze, and social influence in immersive virtual environments. Hum Commun Res 31(4):511–537
113. Bull R, Gibson-Robinson E (1981) The influences of eye-gaze, style of dress, and locality on the amounts of money donated to a charity. Hum Relat 34(10):895–905

114. Baylor AL, Kim Y (2004) Pedagogical agent design: the impact of agent realism, gender, ethnicity, and instructional role. Intelligent Tutoring Systems. Springer, New York, pp 592–603
115. Coe R, Aloisi C, Higgns S, Major LE (2014) What makes great teaching?. Review of the underpinning research. Tech. rep, Sutton Trust
116. Hill HC, Rowan B, Ball DL (2005) Effects of teachers mathematical knowledge for teaching on student achievement. Am Educ Res J 42(2):371–406
117. Askew M, Brown M, Rhodes V, Johnson D, Wiliam D (1997) Effective teachers of numeracy. Kings College, London
118. Garner PW (2010) Emotional competence and its influences on teaching and learning. Educ Psychol Rev 22(3):297–321
119. Ronfeldt M, Loeb S, Wyckoff J (2012) How teacher turnover harms student achievement. Am Educ Res J 50(1):4–36. doi:10.3102/0002831212463813
120. Wang N, Johnson WL, Gratch J (2010) Facial expressions and politeness effect in foreign language training system. In: Intelligent tutoring systems, Springer, pp 165–173
121. Bloom BS (1984) The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. Educ Res 13:4–16
122. VanLehn K (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychol 46(4):197–221
123. Beebe B, Jaffe J, Lachmann F (1992) A dyadic systems view of communication. In: Warshaw S (ed) Relational perspectives in psychoanalysis. Analytic Press, Hillsdale, NJ
124. Jaffe J, Beebe B, Feldstein S, Crown CL, Jasnow MD, Rochat P, Stern DN (2001) Rhythms of dialogue in infancy: coordinated timing in development. Monogr Soc Res Child Dev 66(2):i–149
125. Green J, Weade R (1985) Reading between the words: social cues to lesson participation. Theory Pract 24(1):14–21. doi:10.1080/00405848509543141
126. Nicol D, Minty I, Sinclair C (2003) The social dimensions of online learning. Innov Educ Teach Int 40(3):270–280
127. Baxter P, Wood R, Baroni I, Kennedy J, Nalin M, Belpaeme T (2013) Emergence of turn-taking in unstructured child-robot social interactions. In: Proceedings of the 8th ACM/IEEE international conference on human-robot interaction, IEEE Press, pp 77–78
128. Zajonc RB (1965) Social facilitation. Science 149(3681):269–274

James Kennedy received a B.Sc. (Hons) in Music Systems Engineering and an M.Sc. in Information Technology from the University of the West of England (UK) in 2010 and 2012, respectively. He is currently completing his Ph.D. in Human–Robot Interaction at Plymouth University (UK). His research interests centre around social companion robots, particularly for use in educational interactions with children. He has previously been involved with the EU FP7 ALIZ-E project and is currently working alongside the EU FP7 DREAM project and the EU H2020 L2TOR project.

Paul Baxter is interested in developing and applying adaptive learning robots to social human–robot interaction, particularly in learning and educational contexts. Until recently, he was a Research Fellow in the Centre for Robotics and Neural Systems at Plymouth University (UK), where he worked with the EU FP7 DREAM and ALIZ-E, and H2020 L2TOR projects, focussed on cognitive robots for long-term child-robot social interaction. Prior to this, he completed his Ph.D. at the University of Reading (UK) in the domain of developmental cognitive robotics. He is currently a lecturer in the School of Computer Science at the University of Lincoln (UK).

Tony Belpaeme received his Ph.D. in Computer Science from the Vrije Universiteit Brussel in 2002 and is currently Professor in Robotics and Cognitive Systems at Plymouth University (UK) where he leads a research lab in the Centre for Robotics and Neural Systems. Starting from the premise that cognition is rooted in social interaction, Belpaeme and team try to further the science and technology behind artificial intelligence and social robots. This results in a spectrum of findings, from theoretical insights to practical applications. He coordinated the FP7 ALIZ-E project, and collaborated on the ROBOT-ERA, DREAM and ITALK projects. He is currently the coordinator of the H2020 L2TOR project.

PLOS ONE

RESEARCH ARTICLE

Robot education peers in a situated primary school study: Personalisation promotes child learning

Paul Baxter[1,2]*, Emily Ashurst[2], Robin Read[2], James Kennedy[2], Tony Belpaeme[2,3]

1 Lincoln Centre for Autonomous Systems, School of Computer Science, University of Lincoln, Lincoln, United Kingdom, **2** Centre for Robotics and Neural Systems, The Cognition Institute, Plymouth University, Plymouth, United Kingdom, **3** IDLab – imec, University of Ghent, Ghent, Belgium

* pbaxter@lincoln.ac.uk

Citation: Baxter P, Ashurst E, Read R, Kennedy J, Belpaeme T (2017) Robot education peers in a situated primary school study: Personalisation promotes child learning. PLoS ONE 12(5): e0178126. https://doi.org/10.1371/journal.pone.0178126

Editor: Natalia Reich-Stiebert, CITEC, Bielefeld University, GERMANY

Received: May 5, 2016

Accepted: May 8, 2017

Published: May 23, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the EU FP7 projects ALIZ-E (grant number 248116) and DREAM (grant number 611391, http://dream2020.eu/), and H2020 project L2TOR (grant number 688014, http://www.l2tor.eu). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The benefit of social robots to support child learning in an educational context over an extended period of time is evaluated. Specifically, the effect of personalisation and adaptation of robot social behaviour is assessed. Two autonomous robots were embedded within two matched classrooms of a primary school for a continuous two week period without experimenter supervision to act as learning companions for the children for familiar and novel subjects. Results suggest that while children in both personalised and non-personalised conditions learned, there was increased child learning of a novel subject exhibited when interacting with a robot that personalised its behaviours, with indications that this benefit extended to other class-based performance. Additional evidence was obtained suggesting that there is increased acceptance of the personalised robot peer over a non-personalised version. These results provide the first evidence in support of peer-robot behavioural personalisation having a positive influence on learning when embedded in a learning environment for an extended period of time.

Introduction

Social robots have the potential to make positive contributions to a range of human-centred activities, from support of the elderly to therapeutic assistance to adults and children [1–4]. One domain of particular interest is education, where social robots may be used to supplement existing teaching structures to provide additional support to children. A range of evidence comes together to support this perspective: it is known that one-to-one tutoring leads to significant learning improvements [5], classroom engagement is a predictor for peer acceptance in later years in young children [6], and that personalised social and academic support has been shown to reinforce later achievement [7]. The role of robots to facilitate engagement in classroom activities thus has potentially significant consequences for learning as well as for social development. In these efforts, the role of adaptivity is considered central to the efficacy of application: an adaptive robot will be able to take into account the specific needs, requirements and preferences of the person(s) with whom they are interacting. This personalisation of robot

Competing interests: The authors declare that no competing interests exist.

behaviours is the focus of the present work. In this paper, we demonstrate the positive role that personalised robot peer behaviours play (along a number of dimensions) for child learning in a situated context.

Existing work has shown that the presence of robots confers a number of advantages over other media (e.g. standard desktop computers or paper-based systems) for learning and behavioural change in people [8]. This has been demonstrated, for example, in the domains of adherence to weight-loss programmes [9], reducing puzzle solving times [10], learning words [11], and motor task learning [12]. Further studies have shown that physical robots will attract more attention than their virtual analogues [13–15], and will comply with their requests [16], following evidence suggesting children regard social robots as psychological agents [17] and are perceived as more enjoyable interaction partners [18, 19]. Taken together, these studies indicate that robots take advantage of, and amplify, the human propensity to anthropomorphise inanimate objects, which results in subsequent behavioural change [20, 21]. Given this effect of physical robots as a basis, the question of interest is therefore how the behaviour of the robot can augment this to maximise the desired outcome for the human interactant.

Two prior studies in the domain of social robots for educational contexts have set benchmarks for subsequent research. In the first (single experimental condition) study, a robot was placed in a corridor outside two Japanese classrooms for two weeks (6-7 and 10-11 year-olds, under experimenter supervision), with the nominal task of encouraging the children to learn English in unstructured interactions in break times [22]. This study demonstrated significantly increased vocabulary recall by the children. In the second study, a humanoid robot with a gradually unfolding repertoire of social behaviours was placed within a classroom of 10-11 year-olds in Japan for two months (32 experimental days), although interactions took place outside of normal lesson times and also under constant experimenter supervision [23]. While the examination of learning outcomes for the children was not the focus of the study, with the development of relationships between the children and robot the primary aim, it was shown that children who maintained peer-like interactions with the robot maintained interactions over the extensive experimental period. Extending significantly from these works, the present study focusses explicitly on learning, and being simultaneously embedded both physically and in terms of the curriculum in the classroom itself.

A number of other studies have recently followed from these seminal works to further explore the specific potential role that such social robots can play in helping *children* to *learn*, although typically these have taken place outside of school classrooms or over isolated interaction sessions. While a number of studies demonstrate the benefit of social robots in terms of preference [24] and for adult learning [25], studies with children have shown that personalisation of robot behaviour (e.g. using names) [26] and task content (e.g. increased coverage of subjects in which the children struggle) [27] can lead to modest learning gains in short-term and single interactions, and that collaborative learning between children is facilitated [28]. However, these studies are ambiguous regarding the actual impact of social behaviour on child learning: the presence of robots appears to facilitate increased learning, but the role of social behaviour to extend this effect remains unclear, in contrast to the human-centred theory [29].

In the present work, we specifically examine the role that robot personalisation can play in supporting the learning of children in social interaction with a humanoid robot over longer and more intensive periods of time. We conduct this study within the classrooms themselves, integrated within the school curriculum, and with no experimenters present during proceedings, so as to maximise the ecological validity of our observations, results, and potential utility for real applications. Our findings broadly support the hypothesis that personalisation within interactions facilitates learning.

Situated school study

In an education context, robots could take on a number of social roles, such as tutor or peer, each of which gives rise to certain behavioural expectations. As noted above, both have been found to result in child learning, and both come with the expectation of social behaviour [30, 31]. However, whereas a tutor can be reasonably expected to not make mistakes, there is not necessarily such an expectation for a peer: indeed, it has been found that the robot making mistakes will further encourage child learning [32]. A robot with a more cooperative interaction style has been found to elicit higher levels of engagement when interacting with children [33]. Finally, in terms of preferences, it has been shown that in comparison with a tutor, a peer role is preferred [34]: in the domain of robot companions for diabetic children for example, the robot playing the role of a peer appears to be preferred over a tutor [4, 35]. For the present study, we therefore focus on the role of social robot as peer; a learning companion.

This focus on the peer role entails a greater emphasis on collaborative (involving multiple parties attempt to learn something together [36]) rather than didactic (in the manner of a teacher) interactions between the child and robot. Technology is broadly being highlighted as a means of ameliorating this [37]: child-child interaction studies have shown that collaborations are more effective with jointly visual and manipulable objects [38]. The touchscreen-based task environment we use takes advantage of this effect by implicitly constraining the content of the interaction to the task [39], thus encouraging collaboration and participation (active learning) [40] in a shared task space. It has been previously shown how such a task environment provides an engaging context for child-robot interactions [15, 41, 42].

Our application context is a primary school classroom, with the intent that the robots act autonomously whilst embedded within them. We seek to achieve ecological validity for the study [43]: we emphasise that the robots are not under experimenter supervision during the experiment (the teacher themselves provide this) and thus also not whilst the children interact with the robot, as this detracts from relevance to potential deployment scenarios. Furthermore, we consider the robot to be embedded within the classroom, both in terms of physical presence (in the classroom, and in operation during lesson time), but also in terms of the incorporation of learning material from the children's curriculum. These two points (embeddedness and unsupervised operation) constitute novel extensions to studies in the existing literature.

These considerations contextualise the broad hypothesis of the present study: *that personalisation in a robot learning peer will lead to greater learning effects for children in an embedded educational context*. Four aspects of this broad hypothesis require specification. Firstly, we hold learning to incorporate generalisation in addition to memorisation, following a revision of Bloom's taxonomy [44], which identifies cognitive processes (from remembering to creation) as well as knowledge (from factual to meta-cognitive) as essential educational objectives. Our learning evaluation thus specifically incorporates aspects of application of knowledge to a new context. Secondly, we note that there are a range of potential targets for learning for the children in their educational environment. For this reason, we examine both topics that are part of their existing curriculum (familiar subjects), and ones that are not (novel subjects). Thirdly, the novelty of our classroom-embedded application necessitates an examination of the attitudes of the children in addition to their performance, to begin to assess the wider implications of such an application. We thus attempt to characterise the wider experience of the children over the experimental period. The fourth aspect is the nature and extent of robot behaviour personalisation, which has been stated as ". . .reflect[ing] the needs and requirements of the (social) environment where the robot is operating in" [45] (p20). Consistent with this definition, Lee et al [24] describe three non-exclusive means of increasing robot personalisation that include aspects of behaviour that are not related directly to adaptation per se, but

also to the creation of a *personable* character: increasing friendliness, alteration to fit user preferences, and adaptation over repeated encounters. This indicates a broad and integrated perspective on personalisation; a position that we here subscribe to.

A range of evidence in HRI studies, grounded in multiple other disciplines, may be brought together to further support this perspective for the present work. Mapping onto the definition and characterisation of behaviour personalisation discussed in the previous paragraph, we identify three particular facets of personalisation that are particularly relevant to our task context: adaptation of non-verbal behaviour, personable language content, and alignment to task performance. These encompass both adaptive (non-verbal and task performance adaptation) and personable (language content) behaviours that match the social interaction context (repeated peer-peer interactions in an education setting). Following the phenomenon that humans align their actions to one another, such as linguistic content [46], non-verbal behaviour adaptation follows from and encompasses those aspects of the robot behaviour that are manipulable based on observation of the child's behaviour [47], based on the phenomenon that humans will adapt their behaviour to that of a robot [48]. Personable language content refers to the explicit taking into account of the specific person with which the interaction takes place: for the present study, this entails using the interacting child's name during the interaction [26], and using an informal style for instruction and feedback utterances [49]; being personable as opposed to imperative. Finally, performance alignment is the modification of aspects of the task to align them with the performance of the child [25, 50]. In the present study, such performance alignment is employed at two levels: firstly at the task level, where the children could repeat an individual task, and secondly at a behavioural level, where the performance of the robot is aligned with that of the child [47]. The first and third facets of personalisation effectively constitute a memory of prior interactions, which may subsequently be applied to further interactions.

As stated above, we consider these three facets of personalisation together as a single concept [24]. Evidence from a range of sources indicates that the consideration of single modality interaction cues is insufficient to account for human behaviour, and that instead a fundamentally integrated perspective needs to be taken [51]. For example, emotion perception has been found to require conceptual processing, and is thus open to contextual influences (e.g. visual and social) [52]. Furthermore, recent theoretical developments in the domain of social cognition, emphasising contingent behaviours, suggest that the context of the interaction shapes the individual's disposition to engage in interaction, resulting in a difficulty in handling out-of-context cues [53]. Given that the context is at least partly determined by the interaction partner, this further indicates the importance of coherency of context. Human social interactions naturally integrate all these aspects of personalisation, and so we anticipate that such coherency would also be expected of a nominally social robot. Taken together, and as a first truly embedded study of this type, these lines of evidence motivate and justify our decision to maintain the integration of the three facets of personalisation for the present study.

The study described in this paper seeks to address the broad hypothesis by using a two-condition, between-subject experimental design. Two age- and ability-matched groups of 7-8 year-old children in a U.K. primary school form the subject groups. A single robot is deployed in each group in the same room in which the children engage in their daily lessons (Fig 1), during which time individual children interact with the robot. They engage in a collaborative sorting task with the robot on novel (history—the stone age) and familiar (mathematics—times-tables) topics using a large mediating touchscreen [54] (Fig 1(b)). There are no experimenters present during the interactions, which took place over a continuous two-week period. In the "*Personalised*" condition (**P**), the robot personalises its behaviour along the three defined

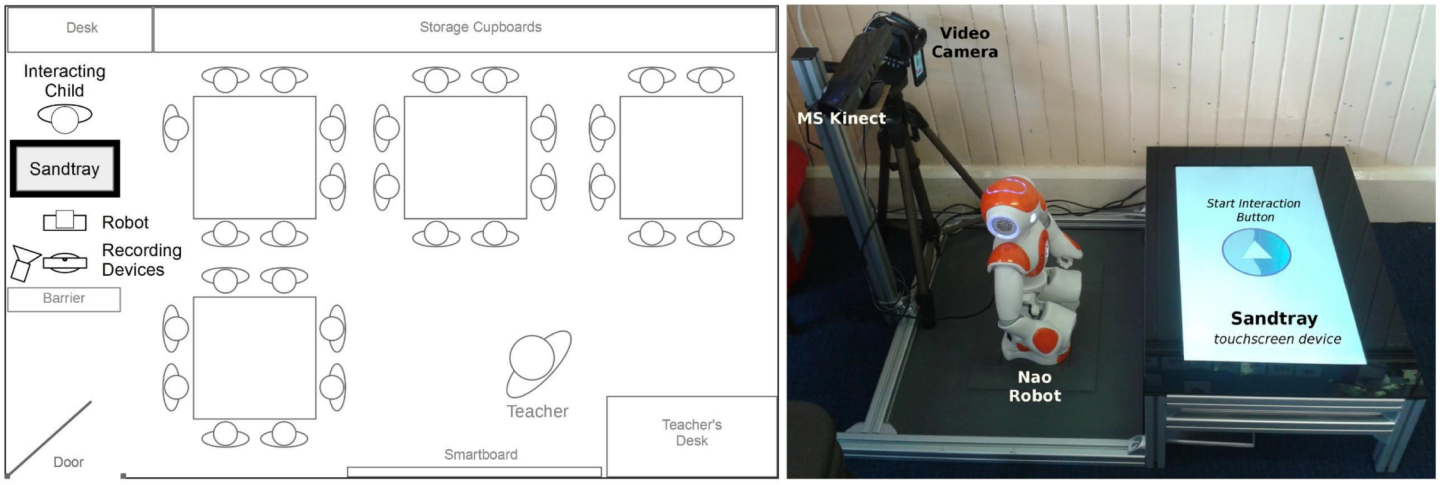


Fig 1. Typical physical setup of the system within the classroom. The robot, Sandtray—a touchscreen device—and camera setup was located in one corner of the room in which the children had their normal lessons. Interactions took place during normal lesson time. Both classrooms had similar arrangements. *Not to scale.*

https://doi.org/10.1371/journal.pone.0178126.g001

dimensions; in the "*Non Personalised*" condition (**NP**) the robot displays non-adaptive, non-personalised behaviour (see the Methods section for details).

Materials and methods

The aim of the study conducted was to investigate whether personalised robots embedded within a classroom for an extended period of time (part of normal classroom activities, and with no experimenters present) can lead to increased child learning. The primary hypothesis of the study is therefore that children in the Personalised robot condition would learn more than children in the Non-Personalised robot condition, on the given set of topics. In addition to this, we seek to explore some of the wider implications of having the robots embedded within the classrooms, and whether the personalisation had any additional effects beyond the target learning outcomes.

Ethics statement

Approval for conducting this study was granted by the Plymouth University Faculty of Science and Technology Human Ethics Committee, as part of a thematic programme of research involving the robot and touchscreen setup, and children in local schools. An opt-out informed consent was obtained in writing from the parents/guardians of all participating children, and a separate opt-in written informed consent was obtained for video recording the interactions between the children and the robots. Children were withdrawn from the study if consent was not obtained, and it was made clear that they could withdraw if and when they wished to.

Subjects

A total of 59 children aged 7-8 (in U.K. year 3) took part in the study (summer term). All children attended a single U.K. primary school, but were divided into two classes. This division

was not on the basis of ability. Gender balance favoured girls, although this applied equally to both the first (12 boys, 18 girls, 30 in total) and second (12 boys, 17 girls, 29 in total) classes.

Each class was based in a different room where the majority of their lessons took place (Information Technology lessons and Sports took place in different areas of the school). These classrooms were located on the same corridor on the first floor of the school building (one other empty classroom was on the same floor). The children in the two classes were separated in these classes, although break times were held in communal areas of the school. Each class was randomly assigned an experimental condition for the duration of the experiment. Each class had a separate teacher who remained with the class for the duration of the experiment period. In addition, each class was assigned a teaching assistant (TA), who varied by day. Both teachers and TAs were briefed regarding the experimental setup; none of these were told of the experimental conditions, nor that there were different robot behaviours deployed in the two classes. This arrangement of children and classes provided the greatest degree of homogeneity possible between the conditions by controlling for a number of potentially confounding subject and environment factors.

Materials

The same hardware setup was employed in both classrooms (Fig 1(a)). This consisted of a touchscreen (the Sandtray), Nao humanoid robot (58cm tall, made by Aldebaran Robotics), aluminium extrusion frame, and recording devices (Fig 1(b)). The robot and touchscreen were synchronised over a wireless network such that the robot could manipulate virtual 'objects' displayed on the screen [54]. The aluminium frame served the dual purpose of maintaining the arrangement of the equipment (e.g. reducing cable trip hazards) and providing a minimal barrier to discourage the children from interfering with the hardware. The only difference between the robots used was the highlight colour of the plastic panels: orange was used in the Personalised condition, and grey was used in the Non-Personalised condition. One such hardware setup was deployed in each classroom, where it remained for the continuous two week period of the experiment.

Learning task

Taking into account the children's current curriculum, two topics for learning in the interaction with the robot were chosen, since there is a suggestion that multiple activities support the maintenance of engagement [55]. The first was *novel* to the children, but was due to be learned in the following academic year. The second was *familiar* as it had already been the ongoing subject of learning. This dual-topic learning task was chosen to assess whether, in the context of a familiar learning environment, a robot learning companion could be applied as an intervention for an existing learning process as well as to a novel task.

The familiar learning task was chosen to be the times-tables, up to and including 12. This formed part of the curriculum that the children studied throughout the year. As such, the children were used to the concept involved, but varied in ability across the subject group. The novel learning task concerned the stone age. This was a new subject matter for the children in the school environment, with it due to appear on the syllabus in the following year. Learning gains made in this topic would thus have been beneficial to the children in the future.

Both topics were administered using the Sandtray, and were structured in the form of a series of two-category sorting tasks played with the robot (e.g. Fig 2(c)). A library of images is placed on the screen, each library comprised of two static category images, and a number of movable images. The task is to sort each movable image into the correct category: visual feedback is displayed on the screen to indicate a correct (or incorrect) categorisation. The child

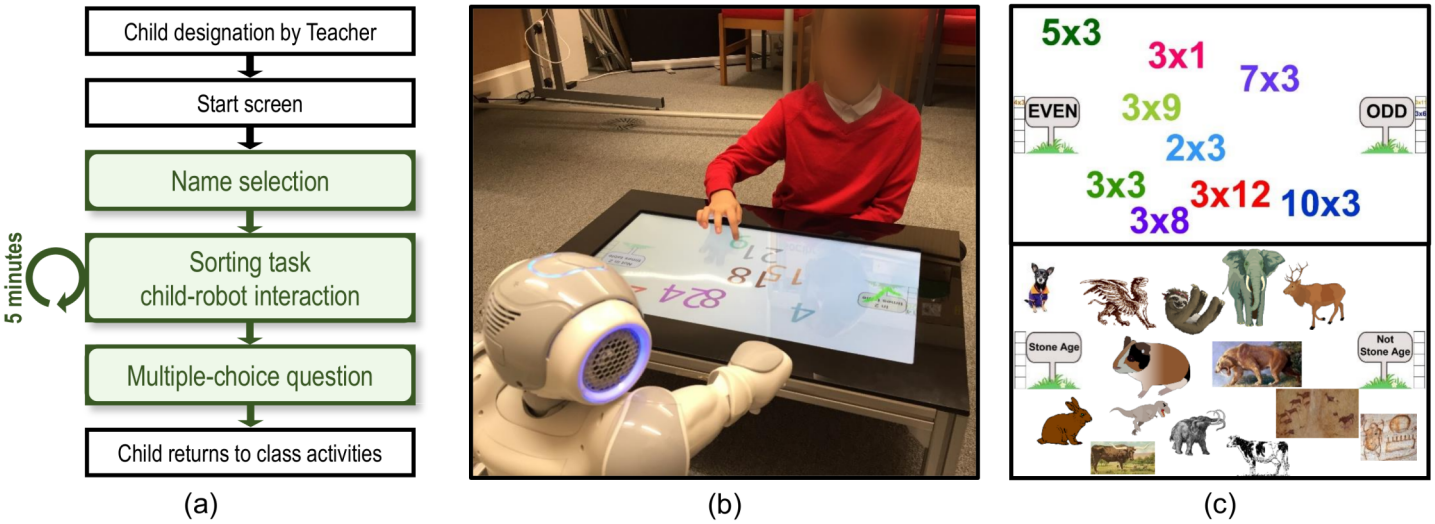


Fig 2. Interaction structure and contents. (a) structure of each interaction, with five minutes on the collaborative sorting task itself; (b) example of a child engaged in the task with the robot (hardware and classroom setup as shown in Fig 1); (c) two sample image libraries, showing a 3 times-table task, and a stone-age animals task.

https://doi.org/10.1371/journal.pone.0178126.g002

uses the touchscreen, and the robot can virtually drag the same images, thus establishing parity of potential interaction affordances with the screen, and facilitating interaction between the child and robot [54]. This methodology has been employed in a number of previous studies [4, 15] and has proven to be an effective strategy to engage children with robot interaction tasks. Given that both novel and familiar learning tasks are displayed on the touchscreen, the tasks are interleaved: i.e. times-tables and stone age libraries are alternated (Table 1).

The image libraries were the same for all children, in both conditions. Each image library formed a two-category sorting task, of which half were uniquely associated with one of the two categories, and half to the other. The stone-age libraries were each comprised of 14 images, and the times-tables libraries were comprised of 12 images. The images appearing in the image libraries did not appear in the pre- and post-experiment knowledge tests. The order of the times-tables is according to difficulty (as specified by the teachers prior to the study), whereas the stone-age image libraries each covers a different topic (where the task is to recognise whether each image displayed belongs in the stone-age or not).

There were two additional learning-related components that were tested in this experiment. In the first, an item of factual information was stated by the robot to the children during their interaction, with recall of this fact tested for at the end of the experiment (with the multiple-choice question "how long ago was the stone-age?", options: {two years, two hundred, two thousand, two million, two trillion, two bazillion}; last option a fake large number, correct answer is two million years ago). The second component was tracking child performance in a class-based task that was independent of either the familiar or novel learning tasks (*incidental* task): spelling test scores were chosen as they were assessed on a weekly basis. In this way, performance prior to, during and after the experiment could be tracked.

Conditions

Two experimental conditions were employed: a Personalised (P) interactive robot condition, and a Non-Personalised (NP) robot condition. The robot behaviour differed between the

Table 1. Image libraries used for the sorting tasks. Shown are the type of sorting task for each library, and the categories used for the sorting itself. There were 14 images per stone age library, and 12 images per times-table library. *Stone age libraries are in italics: the fifth and sixth of these were combinations of images from the first four stone-age libraries.*

Library	Library topic	Library contents	Sorting task
1	Times-table	2x table	In/Out
2	*Stone age*	*Lifestyle*	*Yes/No*
3	Times-table	10x table	In/Out
4	*Stone age*	*Animals*	*Yes/No*
5	Times-table	5x table	Odd/Even
6	*Stone age*	*Tools*	*Yes/No*
7	Times-table	2, 10 & 5 division	Odd/Even
8	*Stone age*	*Art*	*Yes/No*
9	Times-table	3x table	Odd/Even
10	Times-table	4x table	In/Out
11	Times-table	6x table	In/Out
12	Times-table	3, 4, & 6 division	Odd/Even
13	*Stone age*	*mix of subjects*	*Yes/No*
14	Times-table	7x table	In/Out
15	Times-table	8x table	In/Out
16	Times-table	9x table	Odd/Even
17	Times-table	11x & 12x tables	Odd/Even
18	*Stone age*	*mix of subjects*	*Yes/No*

https://doi.org/10.1371/journal.pone.0178126.t001

robots in three distinct respects: non-verbal behaviour (gaze, movement alignment), verbal behaviour (friendliness, personalisation), and adaptivity of progression through the learning content (to personal performance). In neither condition were the children or teachers made aware of the differing aspects of behaviour, nor of the differences between the conditions. In both conditions, the robots acted autonomously, i.e. not under the control of an experimenter or teacher.

In the Personalised condition, the robot was animated (actively seeking to match gazes to it by the interacting child, and exhibiting life-like idling movements), responsive to the approach of a child at the start of an interaction (it would stand up), and varied its behaviour according to the characteristics of each child, as observed in the interaction. In terms of non-verbal behaviour, this constituted adaptation of the drag speed of the robot movements on the screen, the accuracy of the movements (in terms of percentage correct and incorrect categorisations), and the length of time between successive moves [47]. In terms of verbal behaviour, the robot would use the interacting child's name, and employ a more friendly (as opposed to imperative) demeanour. Full details may be found in the supplementary materials (S1 File). Progression through the lesson image libraries was partially dependant on performance: assuming that the child completed more than four image categorisations, then the image library was considered to be successfully completed if the success rate for the child (i.e. not including robot moves) exceeded 65%, with performance below this resulting in the library being repeated (up to a maximum of three times). This personalisation of lesson progress provides a greater degree of opportunity for practice on those topics where performance was low.

For the Non-Personalised condition, the robot's behaviour remained constant throughout all interactions, independent of the characteristics of each child, and was not responsive to the

approach of a child. This included movement speed, accuracy of moves, and delay between moves. Imperative non-personal phrases were used (matched for number and length of utterances used in the Personalised condition), and the progression through the learning material was set at a constant rate for each child: each image library was completed only once before moving on.

In neither condition was there a mechanism to explicitly consider turn-taking behaviours; nevertheless, previous work has indicated that if the children perceive the robot to be a social agent, turn-taking will emerge in the interaction [41].

Protocol

The class teachers were not informed of the hypotheses of the study, nor of the differences in robot behaviour between the classrooms. The teachers administered pre-experiment knowledge tests and questionnaires, and did so again for post-experiment tests and questionnaires. During the experiment period itself, the teachers collected child performance on the normal spelling tests and maths times-table tests, which were administered weekly. Maths lessons were postponed for the two-week duration of the experimental period. A final debriefing interview was conducted with the teachers after the experimental period. These additional data were collected to enable a broader perspective on the influence of the robot in the classroom beyond the interactions themselves.

During the experiment, there were no experimenters in the room: the robot system ran autonomously, with experimenters only present at the start and end of the day to initialise and shut down the system, respectively. In both conditions, the teachers designated the next child to interact with the robot. The child would approach the robot setup (from the right-hand side of Fig 1(b) for example), kneel down, and press a large 'start' button on the screen. Following a verbal acknowledgement from the robot (differing by condition), the child would then proceed to select their name on the screen. On name confirmation, the robot would begin the interaction (differing by condition) with the last uncompleted image library.

After five minutes of interaction time, during which both the child and robot were able to sort the images on the screen, the robot would announce that it had to rest (differing by condition). The child would be asked to answer a multiple-choice question on the screen, the robot would return to it's rest position, and the child would return to their seat in the classroom. The next child could then be called to interact by the teacher.

Metrics

Four types of metric were used: pre- and post-experiment knowledge tests, within-interaction performance data, questionnaires assessing opinion of and engagement with the robot, and measures of performance in the classroom not involved in the experiment.

The pre- and post-experiment knowledge tests were administered on paper on the subject of the novel learning task. They consisted of 24 images, 12 of which belonged to the stone-age category, 12 did not. The same test was administered for both pre and post, but the children were not given any feedback after the pre-test; the images in the test did not appear in the robot interaction stage (Table 1), thereby testing an aspect of generalisation.

Within the interactions, all aspects of the child's performance as detectable by the touchscreen and robot were logged. This included the number of correct and incorrect classification attempts per image library (including repeats in the Personalised condition). The change in performance over interaction time per child could therefore be assessed. In addition to this, at the end of each interaction, the child was asked to answer a multiple-choice question on the screen before returning to their seat in the classroom (Table 2, the precise phrasing depended

Table 2. End-of-interaction questions. Multiple choice questions displayed on the screen after each interaction, each of which had five possible responses.

Int.	Question	Option 1	Option 2	Option 3	Option 4	Option 5
1	Did you enjoy playing?	Not at all	No	A bit	Yes	Yes a lot
2	What would you prefer to play with next?	Robot	Classmates	Read a book	Play outside	Games console
3	What would you prefer to play with next?	Robot	Classmates	Read a book	Play outside	Games console
4+	What do you think of the robot?	Boring	OK	Good	Bad	Brilliant

https://doi.org/10.1371/journal.pone.0178126.t002

on the condition, shown in table S1 File). The questions after interactions two and three were same in order to explore the changes in response over time. The questions varied according to the interaction number, and are shown in Table 2. If the child did not respond within 30 seconds, the interaction would end, and a 'no response' entry was made.

The third type of metric used was the administering of standard questionnaires. A preliminary pre-study questionnaire was administered to provide an indication of prior expectations, following prior work [23]. The main battery of questionnaires was administered after the experiment had been completed. Three questionnaires were used at this time. The first was comprised of two sub-scales of the Intrinsic Motivation Inventory [56, 57]: interest/enjoyment and perceived competence. The second was to assess the perception of social presence of the robot [58], as previously validated [59]. The third was to assess the perceived social support provided by the robot [60], an adaptation of a version validated with children (peer subscale) [61]. All questionnaires may be found in the supplementary materials (S1 File).

The final evaluation metric was performance of the children in a classroom task not related to the topics of the familiar and novel learning tasks. Spelling was determined as a suitable choice for this as it was assessed on a weekly basis, which allowed change in performance to be tracked over the course of the experiment.

Data analysis

For all results, the 95% confidence interval (CI) is provided for both within condition data and between condition comparisons. Where appropriate, normality of data is tested for using the Shapiro-Wilk test [62]; unless otherwise stated, the data are found to be consistent with normality, if not, then the Wilcoxon (non-parametric) test was employed. Homogeneity of data variance is tested for using the Levene's test [63]. Bootstrapping is employed to provide estimations of population hypothesis testing from our collected sample [64]: 10^6 replications are used and the studentized bootstrap 95% CI reported [65].

When considering learning effects, it should be noted that the pre- and post-tests used have a maximum (and minimum) possible score, leading to a negative correlation of absolute learning gain and pre-test score [66]. Given this limit on maximal attainable increase in score, the normalised learning gain metric, $g = (score_{post} - score_{pre})/(score_{max} - score_{pre})$, is employed, which normalises change in score to pre-test score, while being uncorrelated with pre-test score [67]. This enables an assessment of the extent of learning irrespective of prior (starting) performance. Normalised learning gain is calculated for all individuals, with the mean normalised learning gain for each condition subsequently derived (and associated 95% CI).

Results

Two primary aspects of the results are considered. Given the main hypothesis, the effect of the personalisation of robot behaviours on learning outcomes is considered. Then, given the continued presence of the robots in the two classrooms for the two week period, an assessment is

made of how the children's perceptions varied over time, both within and between conditions. All data may be found in the supplmentary materials (S2 File). First however, we summarise the characteristics of the interactions in the two conditions.

Expectations and interaction characteristics

As part of the pre-experiment questionnaires, the expectations of the children were assessed, following [23]. Four questions were asked of the children regarding their perceptions of the robot and how they expected their interactions to be (please refer to S2 File for full wordings and possible responses). The results of this show no effective differences between the two conditions, reinforcing the notion that the subject population is equivalent between conditions. The children generally expected the robot to be like a friend (66.7%, followed by games console, 15.8%, and toy, 10.5%), wanted to know how the robot worked (across conditions, scale 1–5, $M = 4.53$, $n = 59$, 95% CI = [4.34, 4.72]), and wanted to be friends with the robot (across conditions, scale 1–5, $M = 4.71$, $n = 59$, 95% CI = [4.57, 4.85]).

Both robot setups were permanently located in the two classrooms for a two week period. This encompassed nine school days (a school closure occured on one day in the second week). Over the two conditions for the experimental period, a total of 199 interactions took place between the children and the robots—note that each of these took place in the classroom during normal lesson time, and thus other children were present (albeit under the direct supervision of the teacher). Overall, the children completed an average of $M = 4.56$ image libraries ($n = 59$, $SD = 1.10$) per interaction with the robot.

Given the touchscreen-centred nature of the interactions, performance of the individual children on individual image libraries could be recorded and compared between conditions. This progression through the image libraries is shown in Fig 3. In all cases, performance in the Personalised condition exceeds that in the Non-Personalised condition, however, significance is only present in a few of these cases (S2 File). While not a statistically significant effect, note that the difference between the conditions generally increases as progression through the image libraries increase.

Learning outcomes

Three learning topics were considered, and one recall task. The *novel* topic was recognition of stone-age items; the *familiar* topic was the maths times tables (from two to twelve, inclusive); and the *incidental* topic was a weekly spelling test. The recall task was a fact introduced by the robot in its interactions with the children, the memory for which was tested after the experimental period.

The two classes used in this study were not divided on the basis of ability, although they were of the same age. In order to verify that the abilities of the children involved were ability matched with respect to the learning metrics used, we consider the pre-experiment scores in each of the three topics examined. Each of these indicates that the performance is indeed similar in the novel ($M_P = 0.731$, $n_P = 30$, 95% CI = [0.695, 0.766], $M_{NP} = 0.759$, $n_{NP} = 29$, 95% CI = [0.718, 0.799], independent samples two-tailed t-test: $t(57) = 1.097$, $p = .277$), familiar ($M_P = 0.557$, $n_P = 30$, 95% CI = [0.478, 0.635], $M_{NP} = 0.520$, $n_{NP} = 29$, 95% CI = [0.467, 0.574], independent samples two-tailed t-test: $t(57) = 0.821$, $p = .415$) and incidental tasks ($M_P = 0.617$, $n_P = 29$, 95% CI = [0.526, 0.708], $M_{NP} = 0.654$, $n_{NP} = 28$, 95% CI = [0.553, 0.755], independent samples two-tailed t-test: $t(56) = 0.437$, $p = .664$). This justifies the examination of differential learning outcomes in the two conditions.

From the pre-test scores described above, consideration of the post-test scores provides an initial and illustrative indication of the change in performance. For the novel ($M_P = 0.807$, $n_P =$

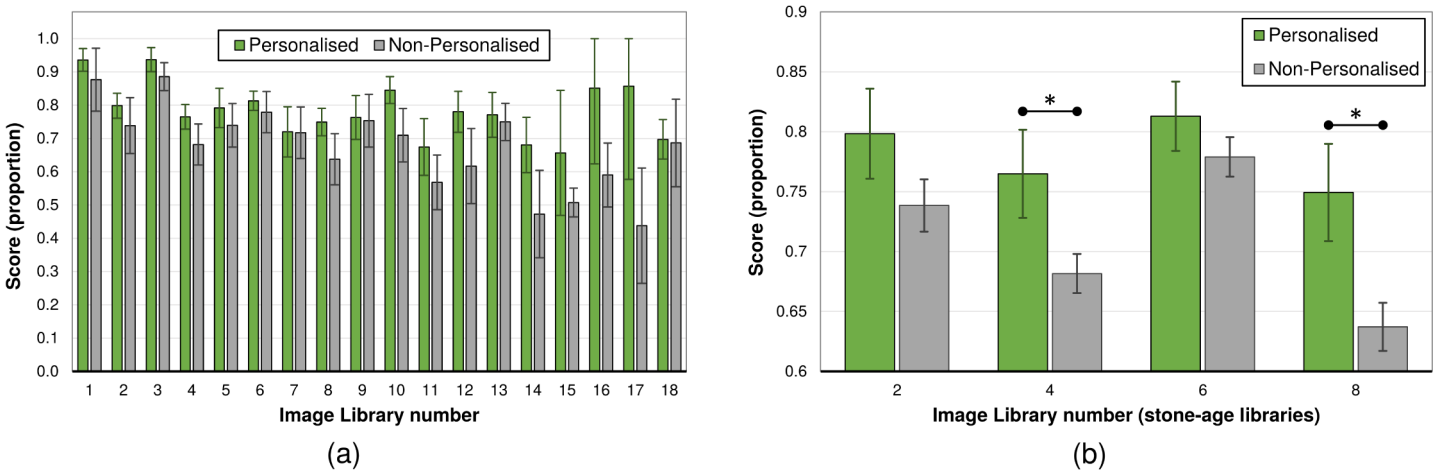


Fig 3. Library scores per image library. Overview of mean scores per library, by condition, error bars are 95% CI: (a) performance in each of the image libraries, see Table 1 for library contents; (b) scores for the first four stone-age image libraries (novel subject): '*' denotes significance at the .05 level.

https://doi.org/10.1371/journal.pone.0178126.g003

30, 95% CI = [0.782, 0.832], M_{NP} = 0.800, n_{NP} = 24, 95% CI = [0.767, 0.834]), familiar (M_P = 0.563, n_P = 30, 95% CI = [0.485, 0.640], M_{NP} = 0.537, n_{NP} = 27, 95% CI = [0.481, 0.592]) and incidental (M_P = 0.800, n_P = 29, 95% CI = [0.697, 0.903], M_{NP} = 0.532, n_{NP} = 28, 95% CI = [0.417, 0.648]) tasks, this indicates similar outcomes between conditions (Fig 4(a)). Only in the incidental task is there an indication of a significant difference between the conditions in the post-test (independent samples two-tailed t-test: $t(55) = 3.396$, $p = .0013$).

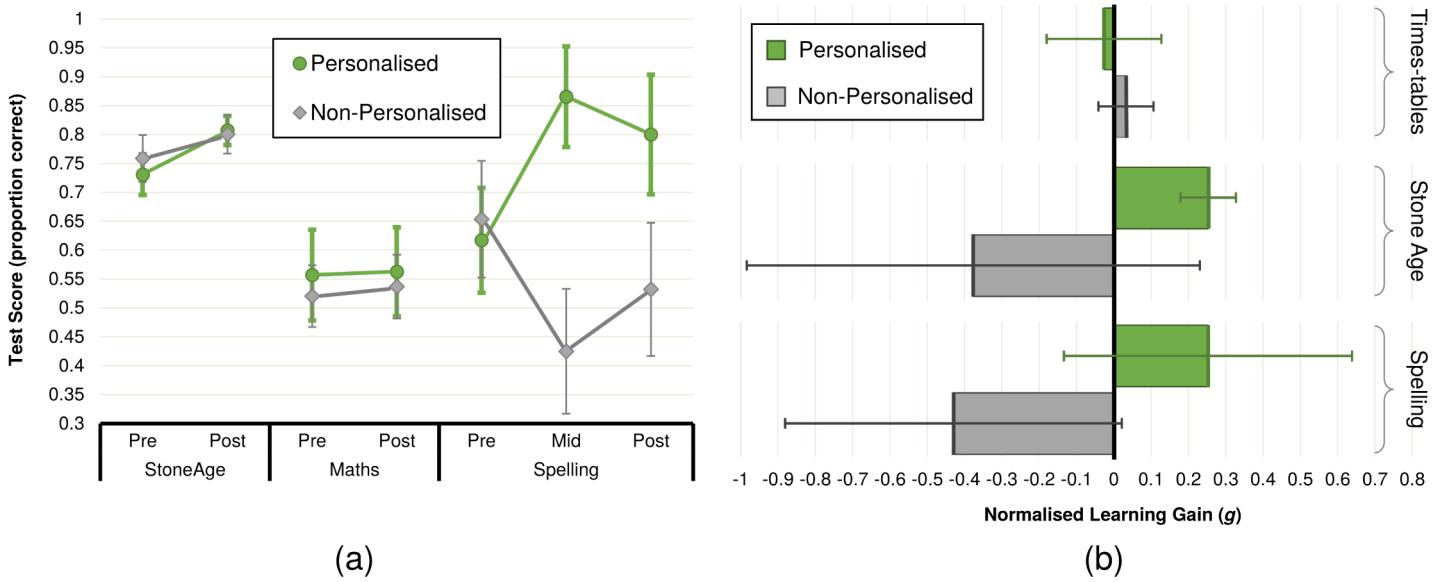


Fig 4. Child learning performance between conditions. (a) summary of mean percentage test scores (for pre and post experimental period) for the familiar learning task (times-tables), the novel learning task (the stone age), and the independent task (spelling, for which there was also a mid-experiment test); (b) normalised learning gain exhibited in the familiar, the novel, and the independent learning tasks. Error bars show 95% CI.

https://doi.org/10.1371/journal.pone.0178126.g004

Table 3. End-of-Interaction Questions Bootstrapping. 10^6 replications on the difference between the conditions (P—NP), compared to observed difference. Numbers in bold denote that observed difference of means lies outside of the bootstrapped 95% CI of the difference of means.

Metric	Difference of the Mean *(P—NP)*	95% CI of bootstrapped difference of means
StoneAge Learning Gain (*novel task*)	**0.629**	[−0.557, 0.589]
Maths Learning Gain (*familiar task*)	−0.059	[−0.174, 0.175]
Spelling Learning Gain (*incidental task*)	**0.682**	[−0.588, 0.589]
Social Presence Questionnaire	0.184	[−0.368, 0.368]
Social Support Questionnaire	0.249	[−0.395, 0.396]
IMI Interest/Enjoyment Questionnaire	0.177	[−0.328, 0.333]
IMI Perceived Competence Questionnaire	0.016	[−0.460, 0.464]

https://doi.org/10.1371/journal.pone.0178126.t003

However, consideration of only the difference between pre- and post-test scores (whether by group or by individual) is a flawed metric since there is a ceiling on the maximum attainable score (100%), and thus also on the maximum attainable increase in score given a pre-test score. To counter this issue, we employ the 'normalised learning gain' metric (see Methods section), which normalises score change to pre-test score. Applied to all subjects in both conditions (i.e. all children in the study, minus exclusions), this indicates no significant learning results for the novel ($M = -0.026$, $n = 54$, 95% CI = [−0.309, 0.256]), familiar ($M = 0.002$, $n = 57$, 95% CI = [−0.085, 0.090]) or incidental ($M = -0.082$, $n = 59$, 95% CI = [−0.379, 0.214]) learning tasks.

Applied on a condition-basis (Fig 4(b)) to the data shows that for the novel task (stone-age) the 95% confidence interval around the observed mean learning gain for the Personalised condition does not include zero ($M_P = 0.253$, $n_P = 30$, 95% CI = [0.179, 0.328]), whereas the Non-Personalised condition does ($M_{NP} = -0.376$, $n_{NP} = 24$, 95% CI = [−0.983, 0.231]). For the familiar ($M_P = -0.026$, $n_P = 30$, 95% CI = [−0.179, 0.128], $M_{NP} = 0.033$, $n_{NP} = 27$, 95% CI = [−0.040, 0.107]) and incidental ($M_P = 0.253$, $n_P = 28$, 95% CI = [−0.133, 0.639], $M_{NP} = 0.429$, $n_{NP} = 27$, 95% CI = [−0.881, 0.022]) tasks, all confidence intervals include zero, indicating that no learning is not an unexpected event (i.e. no significant learning effect).

A bootstrapping process was applied to provide estimations of population hypothesis testing, examining whether the observed difference between the condition means lies outside of the non-parametric bootstrapped distribution (Table 3). The analysis shows that this is the case for the novel ($M_{P-NP} = 0.629$, 95% CI = [−0.557, 0.589]) and the incidental ($M_{P-NP} = 0.682$, 95% CI = [−0.588, 0.589]) learning tasks, indicating positive learning effects in these learning tasks. This is not observed in the familiar learning task ($M_{P-NP} = -0.059$, 95% CI = [−0.174, 0.175]).

The final learning-related metric applied was a recall task. After the second image library (the first stone-age library, see Table 1), the robot would introduce a fact related to the stone-age: how long ago it was. In the experiment post-test (paper-based), a multiple-choice question (six options, see Method section) assessed retention of this fact: correct responses in the P condition (57.1%) exceed those in the NP condition (48.1%), both of which exceed chance (1/6, 16.7%). Application of the Fisher exact test (due to small/null values present in the 6x2 contingency table) reveals a marginal effect ($p = .059$). Collapsing the contingency table into 2x2 (correct/incorrect responses) reveals no significant effect ($\chi^2(2, 55) = 0.446$, $p = .504$). That both condition groups of children perform greater than chance (multinomial probability for both P and NP given 1/6 chance level, $p < .001$) indicates a learning effect. However, given the presence of the robot in the classroom during the interactions, the marginal effect between the conditions could, for example, be due to social contagion effects between individuals of the class.

These results indicate that the interaction with the personalised robot leads to a significantly increased learning outcome for the children in the novel task than with the non-personalised robot, although this is not the case for the familiar task. There is a similar suggestion of increased learning performance for the incidental task, since this was assessed at the same time and in the same way for both condition groups. However, while this result is significant, we only tentatively claim the beneficial role of the personalised robot on other aspects of classroom-based work (as with the familiar task) since there are number of factors for which there was no control put in place (e.g. potential exposure to the material to be learned in the intervening time, or social interaction effects between subjects). This result does however lend significant support to a further exploration of this issue.

Child perceptions and correlations

After each of the interactions, a multiple-choice question was displayed on the screen, with the robot asking the children to choose one of the options prior to returning to their seat (see Table 2). The question posed after the first interaction ("*did you enjoy playing?*") reveals high levels of agreement for both conditions: 96.7% chose "*yes a lot*" or "*yes*" in the personalised condition ($n = 30$), compared with 89.7% in the non-personalised condition ($n = 29$), with no significant difference between the two. This is not a surprising result, given the initial enthusiasm due to the novelty effect.

The questions posed after interactions two and three were the same ("*what would you prefer to play with next?*", with answers classified as either robot or other), and enable an examination of changes in response over time, possibly as the novelty effect increasingly wore off. The results show (Fig 5(a)) that in both conditions there is a reduction in children choosing the robot over other options, with this effect being greater in the NP condition. This difference between interaction numbers is not significant in either the P ($d_{int2-int3} = 0.033$, $\chi^2(2, 60) = 1.355$, $p = .508$) or NP ($d_{int2-int3} = 0.137$, $\chi^2(2, 54) = 2.703$, $p = .259$) conditions. In addition, the effect size is weak for the P condition (Cramer's $V_P = 0.150$), and moderate for the NP condition (Cramer's $V_{NP} = 0.224$). These results suggest that the novelty effect was reducing over the course of the interactions.

The post-experiment questionnaires assessed four aspects of the children's perceptions of the robot: social presence (SPQ), social support (SSQ), interest and enjoyment, and perceived competence; please refer to the supplementary materials for full details of the questionnaires (S1 File). Overall questionnaire reliability (Cronbach's α) was high (listwise deletion for missing values) for the SPQ ($\alpha = 0.878$), SSQ ($\alpha = 0.899$), interest and enjoyment ($\alpha = 0.817$), and for the perceived competence ($\alpha = 0.812$), which indicates good internal consistency.

Overall, the robot was rated highly in terms of social support, the children expressed high levels of interest and enjoyment in the activity and in their own competence, with slightly lower levels of perceived social presence for the robot.

There are however no significant differences between the conditions for any of the four questionnaire-based results: SPQ ($M_P = 3.783$, $n_P = 28$, 95% CI = [3.545, 4.022], $M_{NP} = 3.599$, $n_{NP} = 26$, 95% CI = [3.311, 3.887], independent samples two-tailed t-test: $t(50) = 0.965$, $p = .339$), SSQ ($M_P = 4.247$, $n_P = 28$, 95% CI = [4.012, 4.482], $M_{NP} = 3.998$, $n_{NP} = 26$, 95% CI = [3.673, 4.323], independent samples two-tailed t-test: $t(46) = 1.215$, $p = .231$), Enjoyment/Interest ($M_P = 4.648$, $n_P = 28$, 95% CI = [4.411, 4.884], $M_{NP} = 4.470$, $n_{NP} = 26$, 95% CI = [4.239, 4.702], independent samples two-tailed t-test: $t(52) = 1.051$, $p = .298$), or Competence ($M_P = 4.125$, $n_P = 28$, 95% CI = [3.785, 4.465], $M_{NP} = 4.109$, $n_{NP} = 23$, 95% CI = [3.795, 4.423], independent samples two-tailed t-test: $t(49) = 0.069$, $p = .945$). Bootstrapping supports this by

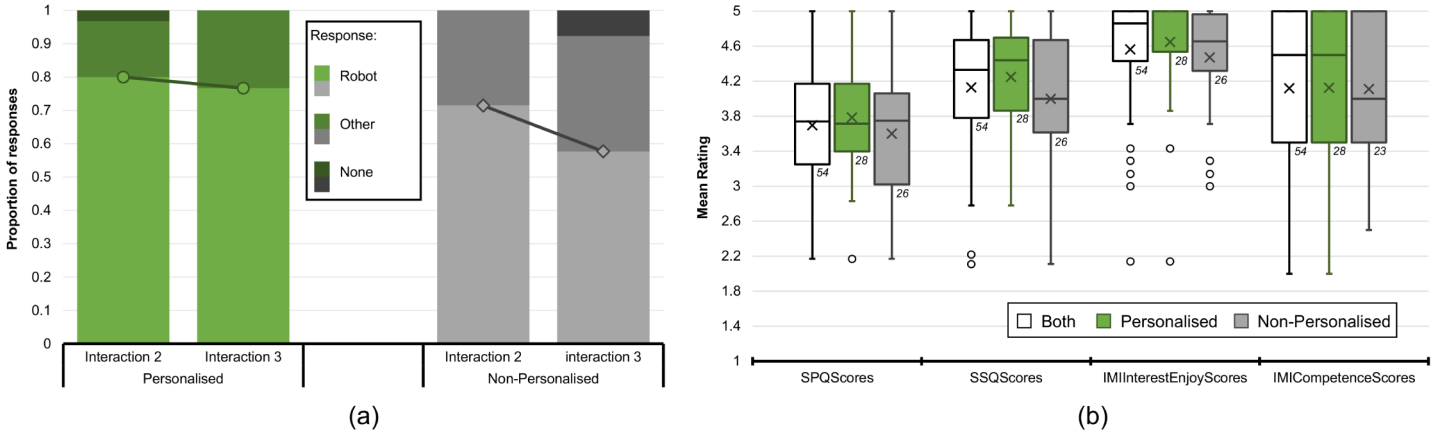


Fig 5. End-of-interaction question responses. (a) end of interaction responses after the second and third interactions to the question "what would you prefer to play with next?", with "none" recorded if an answer is not given within 30 seconds (multiple choice from: robot, classmates, read a book, play outside, games console, or no answer); (b) box-plots showing child ratings for the four questionnaires (end of bars represent last datum within the 1.5*IQR; circles denote outside values; no outliers): social presence, social support, interest/enjoyment and perceived competence. Crosses indicate the mean, numbers below the bars denote sample size.

https://doi.org/10.1371/journal.pone.0178126.g005

showing a lack of significant difference between the conditions with respect to these four aspects of robot perception (Table 3).

It is also of interest to examine the relationship between the performance levels, responses, and questionnaire answers. Correlations are used for this (as opposed to linear regression) since all variables are measured rather than manipulated (except for the conditions themselves): we seek to explore the data rather than generate predictions. The majority of correlations are not significant, or are the same in both conditions. However, a number of observations can be made based on the significance (or not) of the correlations in both the P (Table 4) and NP (Table 5) conditions. In the NP condition, the score attained in the first interaction is strongly and positively correlated with the first question response (whether they enjoyed the interaction: $r(26) = 0.542$, $p = .003$), whereas this is not the case for the P condition ($r(28) = 0.097$, $p = .610$), despite the mean scores ($M_P = 0.798$, $M_{NP} = 0.756$) and responses ($M_P = 2.867$, $M_{NP} = 2.643$) being equally high. Conversely, however, the response in

Table 4. P-condition correlations. Pearson product-moment correlation coefficients for the P condition between the post-experiment questionnaires, first interaction score and response, and the overall learning gain. Cells in bold denote correlations significant at least at the .05 level.

	SPQ	SSQ	Int / Enj	Comp	Int1 score	Int1 resp	SA-gain	M-gain	S-gain
SPQ	1								
SSQ	**0.675**	1							
Int/Enj	**0.466**	**0.518**	1						
Comp	**0.467**	**0.378**	**0.498**	1					
Int1 score	0.094	0.042	0.026	−0.108	1				
Int1 resp	0.251	0.214	**0.743**	0.359	0.097	1			
SA-gain	0.175	−0.076	−0.208	−0.011	0.327	−0.095	1		
M-gain	−0.151	0.159	0.138	−0.234	−0.172	0.083	−0.250	1	
S-gain	−0.189	0.079	0.253	−0.102	0.127	−0.066	−0.344	0.065	1

https://doi.org/10.1371/journal.pone.0178126.t004

Table 5. NP-condition correlations. Pearson product-moment correlation coefficients for the NP condition between the post-experiment questionnaires, first interaction score and response, and the overall learning gain. Cells in bold denote correlations significant at least at the .05 level.

	SPQ	SSQ	Int / Enj	Comp	Int1 score	Int1 resp	SA-gain	M-gain	S-gain
SPQ	1								
SSQ	**0.748**	1							
Int/Enj	**0.443**	0.335	1						
Comp	0.400	0.363	0.101	1					
Int 1 score	−0.074	0.046	−0.224	−0.307	1				
Int 1 resp	0.049	0.014	−0.032	0.142	**0.542**	1			
SA-gain	0.271	0.228	0.126	−0.001	0.079	0.207	1		
M-gain	−0.089	0.124	0.326	0.017	−0.135	−0.077	−0.157	1	
S-gain	**−0.476**	−0.311	−0.272	0.071	0.262	0.243	−0.057	0.097	1

https://doi.org/10.1371/journal.pone.0178126.t005

interaction one is strongly and positively correlated with the interest/enjoyment post-experiment questionnaire response in the P condition ($r(26) = 0.743$, $p<.001$), but not in the NP condition ($r(24) = -0.032$, $p = .877$). This appears to suggest that the levels of enjoyment experienced in the first interaction are maintained throughout the experiment in the P condition, but not necessarily in the NP condition. The correlations between the post-experiment questionnaire responses are similar between the two conditions, with the exception of a significant positive correlation between perceived competence and interest/enjoyment for the P condition ($r(26) = 0.498$, $p = .001$), but not for the NP condition ($r(21) = 0.101$, $p = .655$).

Taken together, these results indicate a high level of continued engagement with the robot is sustained in both conditions, even after the two-week experimental period. There is some indication that, where this existed in the first place, this is sustained somewhat more in the P condition than in the NP condition.

Discussion and conclusion

In general terms, the results show that children exhibit significantly increased learning in the novel learning task in the personalised condition compared with the non-personalised condition. This effect is also apparent in the incidental learning task, but not in the familiar learning task. Personalisation encompasses three distinct aspects (non-verbal behaviour, linguistic content, and performance alignment) that we consider as contributing to the integrated perception of a single agent: in addition to the cue integration framework [51], discontinuities between different aspects of the robot behaviour (e.g. personalisation in one respect, but not in another) may impair the overall perception [68]. This motivated our decision to provide the comparison between an integrated personalisation agent and one that did not, with the subsequently observed differences in learning outcome.

One aspect of the results that may have been impacted by this amalgamation of features in the implementation of personalisation is the perceived 'friendliness' of the robot, which has been characterised as including gentle, predictable movements [69]. It is thus possible that the difference in robot personalisation between conditions leads to a difference in perception of friendliness, which in turn could have an effect on the learning outcomes. However, the outcome of the post-study questionnaires indicates that that this is not the case. Specifically, the Social Presence (SPQ), Social Support (SSQ), and interest and enjoyment questionnaires all showed non-significant differences between the conditions. To the extent that the SPQ and

SSQ responses are related to friendliness, this indicates that friendliness is not a confounding factor for the learning results.

In terms of behaviour, two further characteristics in particular can be incorporated beyond the three aspects of personalisation currently used, namely personality and affective responsiveness. Adaptation to personality has, with adults for example, been shown to be beneficial in the domains of the home [70], rehabilitation [71], and human-robot collaboration [72]. The incorporation of such adaptation for children in an educational context may thus be of interest in the future, even if the reliability of child self-report personality assessments may be questionable [73]. Affective responsiveness for a robot, as a more reactive phenomenon, has been associated with a greater perception of social support [60], with the face of the robot a particularly important feature [74]. A limitation in the current study regards the expressivity of the hardware platform, particularly in terms of variation in facial expression (the Nao robot used has a minimal static face, see Fig 1(b)), which limited the degree to which affective responsiveness, and hence potential for engagement [75], could be achieved. However, the present study nevertheless provides a foundation for further investigation into such issues, by establishing the importance of personalisation for learning.

The embedded nature of the present study methodology contributes to its novelty: we wish to reiterate that the robots became permanent fixtures in the two classrooms over the two week experimental period, and that there were no experimenters/technicians present with the robots during the school day. This remains a rarity in social robotics research. With only the teacher (and occasionally a teaching assistant) present with the children in the classroom, this enabled us to approximate 'natural' conditions for the experiment, thus supporting the ecological validity of our results. There is necessarily however a trade-off for the levels of control over potential influences in an experimental sense [43]. For example, we did not, and indeed could not given the lack of experimenter present, prevent the interaction of individual children with their classmates during their turn with the robot. Furthermore, given that the children of the two separate classes had breaks at the same time, we cannot exclude the possibility that the two groups did not exchange ideas regarding the robot and its behaviour.

The lack of significance between conditions in the familiar task may be due to four effects, apart from the possibility that there are no actual differences to be found. Firstly, robot personalisation as instantiated in the present study may not be sufficient to give rise to outcome differences, or the robot personalisation aspects used were insufficient. However, given the learning differences seen for the novel learning material, we suggest that this is not the case. We certainly acknowledge the possibility of further behavioural refinements, but the demonstration of significantly different learning gains supports our primary hypothesis. Secondly, it is possible that the novelty factor of having robots in the classroom increased overall motivation and hence performance in the tasks. This is unlikely for two reasons: (a) given the same hardware setup in both classrooms, there is nevertheless an increased performance in the novel task for the personalised condition but not the non-personalised condition, indicating the influence of condition differences over a novelty factor; and (b) the qualitative results indicate that the novelty factor decreased in the second week (also see point below). Thirdly, given the potential mixing of children between the conditions outside of the classroom as noted above, there is a possibility of some degree of cross-condition contamination. Whilst the prevalence of this is not possible to rule out, we note in mitigation that the teachers in their debriefings did not suggest that this occurred. We further note that our efforts to maximise the ecological validity of the study necessarily prevented an explicit control for the possible presence of this phenomena. Finally, we recognise that there are limitations in the administration of questionnaires to children, in terms of the ceiling effect, or social desirability distortion [76].

Although our use of standardised questionnaires mitigates the impact of this, the effect remains potentially apparent in the results (Fig 5(b)).

Nevertheless, the experimental design (developed in conjunction with the teachers themselves) sought to avoid and minimise any potential confounds. For example, the teachers were not informed of the specific hypotheses nor conditions of the study, and were involved only in the learning task content and procedural issues (to ensure that similar methods would be used by the teachers when interacting with and referring to the robot in their classroom). Similarly, the classes were balanced in terms of age, gender and ability (as evidenced by the lack of significant difference in the pre-experiment scores and attitudes), reinforced by equivalent pre-experiment expectations, resulting in homogeneous condition groups, which validates our results and observations [77].

The children who took part in the study were primary school children, an age range that has recently seen increasing use in HRI studies [4, 27, 50, 60], as means of supplementing existing educational practice [29]. In terms of generalising the results to other children of the same age, the UK government Office for Standards in Education, Children's Services and Skills (Ofsted) conducts regular school inspections and compiles national statistics and performance tables [78]. For the school at which this study was conducted, the proportion of children who attained the expected standard in reading, writing and mathematics (72%, 2014 rating) for the age group (Key Stage 2, level 4) is consistent with the regional (74%) and national (78%) mean ratings. Based on this characterisation, we suggest that the results could be reasonably generalised to other primary school populations (at least in the U.K.), thus supporting the wider applicability of the findings.

One further point of note is the wider effect of the presence of the robots in the classroom. The teacher debriefing highlighted the impact of novelty: in the first week of the experiment, some disruption to the class occurred as children were distracted by the robot actions and speech. However, they noted that in both classrooms, this distracting effect dissipated in the second week, although they reported still being able to use the robots as a motivator for the children [79]. This is supported by the high levels of interest/enjoyment in the activity at the end of the study (non-significantly higher for the personalised condition). This maintenance of motivation speaks to the wider role of technology, including social robotics, in the classroom and how it is handled ('orchestrated') by the teachers [80]. While acceptance was high in the present study, this may be a self-selection bias (i.e. the school and teachers were enthusiastic about the study prior to implementation), and further examination of the effort required on the part of the teachers and the school versus the learning benefits afforded by the type of personalised social robot systems we have demonstrated here is necessary, particularly in embedded applications (i.e. inside the classroom itself), as we have achieved in the present study.

The methodology employed, with the autonomous robots embedded (both physically and in terms of curriculum) within primary school classes without experimenter supervision, maximises the ecological validity of the study, and thus the implications for educational practice and application. This study found that a robot peer exhibiting personalised behaviours in a collaborative learning task with individual children facilitated improved learning for the children in a novel task over a non-personalised robot behaviour. This effect was not seen for the familiar task, and whilst a differential improvement was observed in the incidental task, these results require further verification in light of the non-significant differences between the child perceptions. We conclude that while further empirical study is required to distinguish between, and indeed maximise the impact of, the different aspects of personalisation employed, we have shown that robot personalisation provides a positive influence on child learning in the classroom.

Supporting information

S1 File. Robot behaviour details and questionnaires. Full details of the robot behaviour in the two experimental conditions, and transcript of the questionnaires used in the post-study debriefing.
(PDF)

S2 File. Experimental data spreadsheet. All details of the experimental data, organised by spreadsheet tab, including pre/post and within-interaction learning and performance data, and screen question and questionnaire responses.
(ZIP)

Acknowledgments

The authors wish to thank Salisbury Road Primary school (Plymouth, U.K.) for their participation.

Author Contributions

Conceptualization: PB TB.

Data curation: PB EA.

Formal analysis: EA PB.

Funding acquisition: TB.

Investigation: EA PB RR JK.

Methodology: PB EA TB RR JK.

Project administration: TB PB.

Resources: TB.

Software: PB RR JK.

Supervision: TB PB.

Validation: PB EA JK RR TB.

Visualization: PB.

Writing – original draft: PB.

Writing – review & editing: PB TB JK.

References

1. Dautenhahn K, Werry I. Towards Interactive Robots in Autism Therapy: Background, Motivation and Challenges. Pragmatics and Cognition. 2004; 12(1):1–35. https://doi.org/10.1075/pc.12.1.03dau
2. Tapus A, Mataric MJ, Scassellati B. The Grand Challenges in Socially Assistive Robotics. IEEE Robotics and Automation Magazine. 2007; 14(1):35–42. https://doi.org/10.1109/MRA.2007.339605
3. Broadbent E, Stafford R, MacDonald B. Acceptance of healthcare robots for the older population: Review and future directions. International Journal of Social Robotics. 2009; 1(4):319–330. https://doi.org/10.1007/s12369-009-0030-6
4. Belpaeme T, Baxter P, Read R, Wood R, Cuayahuitl H, Kiefer B, et al. Multimodal Child-Robot Interaction: Building Social Bonds. Journal of Human-Robot Interaction. 2012; 1(2):33–53.

5. Bloom BS. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational Researcher. 1984; 13(6):4–16. https://doi.org/10.3102/0013189X013006004
6. Hughes JN, Kwok OM. Classroom engagement mediates the effect of teacher-student support on elementary students' peer acceptance: A prospective analysis. Journal of School Psychology. 2006; 43 (6):465–480. https://doi.org/10.1016/j.jsp.2005.10.001
7. Klem AM, Connell JP. Relationships matter: linking teacher support to student engagement and achievement. The Journal of School Health. 2004; 74(7):262–273. https://doi.org/10.1111/j.1746-1561.2004.tb08283.x
8. Li J. The benefit of being socially present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. International Journal of Human Computer Studies. 2015; 77:23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001
9. Kidd CD, Breazeal C. Robots at Home: Understanding Long-Term Human-Robot Interaction. In: Intelligent Robots and Systems (IROS). Nice, France: IEEE Press; 2008. p. 22–26.
10. Leyzberg D, Spaulding S, Toneva M, Scassellati B. The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. In: 34th Annual Conference of the Cognitive Science Society. 1. Sapporo, Japan; 2012. p. 1882–1887.
11. Moriguchi Y, Kanda T, Ishiguro H, Shimada Y, Itakura S. Can young children learn words from a robot? Interaction Studies. 2011; 12(1):107–118. https://doi.org/10.1075/is.12.1.04mor
12. Fridin M, Belokopytov M. Embodied robot vs. virtual agent: involvement of pre-school children in motor task performance. International Journal of Human-Computer Interaction. 2014; 30(6):459–469. https://doi.org/10.1080/10447318.2014.888500
13. Wainer J, Feil-seifer DJ, Shell DA, Mataric MJ. The role of physical embodiment in human-robot interaction. In: IEEE International Symposium on Robot and Human Interactive Communication—RO-MAN. Hatfield, U.K.: IEEE Press; 2006. p. 117–122.
14. Looije R, Van Der Zalm A, Neerincx Ma, Beun RJ. Help, I need some body: the effect of embodiment on playful learning. In: IEEE International Workshop on Robot and Human Interactive Communication; 2012. p. 718–724.
15. Kennedy J, Baxter P, Belpaeme T. Comparing Robot Embodiments in a Guided Discovery Learning Interaction with Children. International Journal of Social Robotics. 2015; 7(2):293–308. https://doi.org/10.1007/s12369-014-0277-4
16. Kennedy J, Baxter P, Belpaeme T. Children comply with a robot's indirect requests. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction—HRI'14. Bielefeld, Germany: ACM Press; 2014. p. 198–199.
17. Meltzoff AN, Brooks R, Shon AP, Rao RPN. "Social" robots are psychological agents for infants: A test of gaze following. Neural Networks. 2010; 23(8-9):966–972. https://doi.org/10.1016/j.neunet.2010.09.005
18. Wainer J, Feil-Seifer DJ, Shell DA, Matari MJ. Embodiment and Human-Robot Interaction: A Task-Based Perspective. In: 16th IEEE International Conference on Robot & Human Interactive Communication. Jeju, Korea: IEEE Press; 2007. p. 872–877.
19. Kose-Bagci H, Ferrari E, Dautenhahn K, Syrdal DS, Nehaniv CL. Effects of Embodiment and Gestures on Social Interaction in Drumming Games with a Humanoid Robot. Advanced Robotics. 2009; 23 (14):1951–1996. https://doi.org/10.1163/016918609X12518783330360
20. Reeves B, Nass C. The Media Equation. Cambridge, MA: Cambridge University Press; 1996.
21. Duffy BR. Anthropomorphism and the social robot. Robotics and Autonomous Systems. 2003; 42:177–190. https://doi.org/10.1016/S0921-8890(02)00374-3
22. Kanda T, Hirano T, Eaton D, Ishiguro H. Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. Human-Computer Interaction. 2004; 19(1):61–84.
23. Kanda T, Sato R, Saiwaki N, Ishiguro H. A two-month field trial in an elementary school for long-term human-robot interaction. IEEE Transactions on Robotics. 2007; 23(5):962–971. https://doi.org/10.1109/TRO.2007.904904
24. Lee M, Forlizzi J, Kiesler S. Personalization in HRI: A longitudinal field experiment. In: HRI 2012; 2012. p. 319–326.
25. Leyzberg D, Spaulding S, Scassellati B. Personalizing robot tutors to individuals' learning differences. Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction—HRI'14. 2014; p. 423–430.
26. Blanson Henkemans O, Bierman BPB, Janssen J, Neerincx M, Looije R, van der Bosch H, et al. Using a robot to personalise health education for children with diabetes type 1: a pilot study. Patient Education and Counseling. 2013; 92(2):174–81. https://doi.org/10.1016/j.pec.2013.04.012

27. Kennedy J, Baxter P, Belpaeme T. The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In: Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction. Portland, Oregon, USA: ACM Press; 2015. p. 67–74.
28. Kanda T, Shimada M, Koizumi S. Children learning with a social robot. In: Proceedings of the 7th ACM/IEEE international conference on Human-Robot Interaction—HRI'12. Boston, MA, U.S.A.: ACM Press; 2012. p. 351–358.
29. Meltzoff AN, Kuhl PK, Movellan J, Sejnowski T. Foundations for a New Science of Learning. Science. 2009; 325(July):284–289. https://doi.org/10.1126/science.1175626
30. Kirby R, Forlizzi J, Simmons R. Affective social robots. Robotics and Autonomous Systems. 2010; 58 (3):322–332. https://doi.org/10.1016/j.robot.2009.09.015
31. Saerbeck M, Schut T, Bartneck C, Janse MD. Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor. In: CHI 2010. Atlanta, USA: ACM Press; 2010. p. 1613–1622.
32. Tanaka F, Matsuzoe S. Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning. Journal of Human-Robot Interaction. 2012; 1 (1):78–95. https://doi.org/10.5898/JHRI.1.1.Tanaka
33. Okita SY, Ng-Thow-Hing V, Sarvadevabhatla RK. Multimodal approach to affective human-robot interaction design with children. ACM Transactions on Interactive Intelligent Systems. 2011; 1(1):1–29. https://doi.org/10.1145/2030365.2030370
34. Looije R, Neerincx MA, de Lange V. Children's responses and opinion on three bots that motivate, educate and play. Journal of Physical Agents. 2008; 2(2):13–20.
35. Baroni I, Nalin M, Baxter P, Pozzi C, Oleari E, Sanna A, et al. What a robotic companion could do for a diabetic child. In: The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RoMAN'14). Edinburgh, U.K.: IEEE Press; 2014. p. 936–941.
36. Dillenbourg P. What do you mean by collaborative learning? In: Dillenbourg P, editor. Collaborative Learning: Cognitive and Computational Approaches. Elsevier; 1999. p. 1–15.
37. Crook C. Children as computer users: the case of collaborative learning. Computers & Education. 1998; 30(3-4):237–247. https://doi.org/10.1016/S0360-1315(97)00067-5
38. Verba M. The Beginnings of Collaboration in Peer Interaction. Human Development. 1994; 37(3):125–139. https://doi.org/10.1159/000278249
39. Kennedy J, Baxter P, Belpaeme T. Constraining Content in Mediated Unstructured Social Interactions: Studies in the Wild. In: 5th International Workshop on Affective Interaction in Natural Environments at ACII 2013. Geneva, Switzerland: IEEE Press; 2013. p. 728–733.
40. Prince M. Does Active Learning Work? A Review of the Research. Journal of Engineering Education. 2004; 93(July):223–231. https://doi.org/10.1002/j.2168-9830.2004.tb00809.x
41. Baxter P, Wood R, Baroni I, Kennedy J, Nalin M, Belpaeme T. Emergence of Turn-taking in Unstructured Child-Robot Social Interactions. In: HRI'13. 1. Tokyo, Japan: ACM Press; 2013. p. 77–78.
42. Hood D, Lemaignan S, Dillenbourg P. When Children Teach a Robot to Write: An Autonomous Teachable Humanoid Which Uses Simulated Handwriting. In: HRI 2015. Portland, OR, USA: ACM Press; 2015.
43. Ros R, Nalin M, Wood R, Baxter P, Looiije R, Demiris Y, et al. Child-Robot Interaction in The Wild: Advice to the Aspiring Experimenter. In: ICMI. Alicante, Spain; 2011. p. 335–342.
44. Krathwohl DR. A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice. 2002; 41(4):212–218. https://doi.org/10.1207/s15430421tip4104_2
45. Dautenhahn K. Robots We Like to Live With?! – A Developmental Perspective on a Personalized, Life-Long Robot Companion. In: ROMAN'04. Kurashiki, Japan: IEEE Press; 2004. p. 17–22.
46. Pickering MJ, Garrod S. Toward a mechanistic psychology of dialogue. The Behavioral and Brain Sciences. 2004; 27(2):169–90; discussion 190–226. https://doi.org/10.1017/S0140525X04000056
47. Baxter PE, de Greeff J, Belpaeme T. Cognitive architecture for human–robot interaction: Towards behavioural alignment. Biologically Inspired Cognitive Architectures. 2013; 6:30–39. https://doi.org/10.1016/j.bica.2013.07.002
48. Vollmer AL, Rohlfing KJ, Wrede B, Cangelosi A. Alignment to the Actions of a Robot. International Journal of Social Robotics. 2014;
49. Cassell J, Bickmore T. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. User Modeling and User-Adapted Interaction. 2003; 13(1-2):89–132. https://doi.org/10.1023/A:1024026532471

50. Janssen JB, van der Wal CC, Neerincx MA, Looije R. Motivating children to learn arithmetic with an adaptive robot game. In: Proceedings of the Third International Conference on Social Robotics. ICSR'11. Berlin, Heidelberg: Springer-Verlag; 2011. p. 153–162.
51. Zaki J. Cue Integration: A Common Framework for Social Cognition and Physical Perception. Perspectives on Psychological Science. 2013; 8(3):296–312. https://doi.org/10.1177/1745691613475454
52. Nook EC, Lindquist Ka, Zaki J. A New Look at Emotion Perception: Concepts Speed and Shape Facial Emotion Recognition. Emotion. 2015; 15(5):569–578. https://doi.org/10.1037/a0039166
53. Di Paolo E, De Jaegher H. The interactive brain hypothesis. Frontiers in Human Neuroscience. 2012; 6 (June):1–16.
54. Baxter P, Wood R, Belpaeme T. A Touchscreen-Based 'Sandtray' to Facilitate, Mediate and Contextualise Human-Robot Social Interaction. In: HRI'12. Boston, MA, U.S.A.; 2012. p. 105–106.
55. Coninx A, Baxter P, Oleari E, Bellini S, Bierman B, Henkemans OB, et al. Towards Long-Term Social Child-Robot Interaction: Using Multi-Activity Switching to Engage Young Users. Journal of Human-Robot Interaction. 2016; 5:32–67. https://doi.org/10.5898/JHRI.5.1.Coninx
56. Ryan RM. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. Journal of Personality and Social Psychology. 1982; 43(3):450–461. https://doi.org/10.1037/0022-3514.43.3.450
57. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist. 2000; 55(1):68–78. https://doi.org/10.1037/0003-066X.55.1.68
58. Leite I, Martinho C, Pereira A, Paiva A. As time goes by: Long-term evaluation of social presence in robotic companions. Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication. 2009; p. 669–674.
59. Harms C, Biocca F. Internal Consistency and Reliability of the Networked Minds Measure of Social Presence. In: Alcaniz M, Rey B, editors. Seventh Annual International Workshop: Presence. Valencia, Spain; 2004.
60. Leite I, Castellano G, Pereira A, Martinho C, Paiva A. Long-term interactions with empathic robots: Evaluating perceived support in children. In: International Conference on Social Robotics. vol. LNAI 7621. Chengdu, China: LNCS; 2012. p. 298–307.
61. Gordon AT. Assessing Social Support in Children: Development and Initial Validation of the Social Support Questionnaire for Children. Louisiana State University; 2011.
62. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics. 2011; 2(1):21–33.
63. Levene H. Robust tests for equality of variances. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. 1960; 2:278–292.
64. Hall P, Wilson SR. Two Guidelines for Bootstrap Hypothesis Testing. Biometrics. 1991; 47(2):757–762. https://doi.org/10.2307/2532163
65. Carpenter J, Bithell J. Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. Statistics in Medicine. 2000; 19(9):1141–1164. https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9%3C1141::AID-SIM479%3E3.0.CO;2-F
66. Hake RR. Interactive Engagement vs Traditional Methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Americal Journal of Physics. 1998; 66(64).
67. Meltzer DE. The relationship between mathematics preparation and conceptual learning gains in physics: A possible hidden variable in diagnostic pretest scores. American Journal of Physics. 2002; 70 (12):1259. https://doi.org/10.1119/1.1514215
68. Koay K, Syrdal D, Dautenhahn K, Arent K, Malek L, Kreczmer B. Companion migration–initial participants' feedback from a video-based prototyping study. In: Mixed Reality and Human-Robot Interaction. Springer; 2011. p. 133–151.
69. Dario P, Guglielmelli E, Laschi C. Humanoids and Personal robots: Design and experiments. Journal of Robotic Systems. 2001; 18(12):673–690. https://doi.org/10.1002/rob.8106
70. Woods S, Dautenhahn K, Kaouri C, Boekhorst RT, Koay KL, Walters ML. Are robots like people?: Relationships between participant and robot personality traits in human–robot interaction studies. Interaction Studies. 2007; 8(2):281–305. https://doi.org/10.1075/is.8.2.06woo
71. Tapus A, Tapus C, Matarić MJ. User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. Intelligent Service Robotics. 2008; 1(2):169–183. https://doi.org/10.1007/s11370-008-0017-4
72. Ivaldi S, Lefort S, Peters J, Chetouani M, Provasi J, Zibetti E. 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. International Journal of Social Robotics. 2016; p. 1–24.

73. Baxter P, Belpaeme T. 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; 2016.

74. Castellano G, Pereira A, Leite I, Paiva A, Mcowan PW. Detecting User Engagement with a Robot Companion Using Task and Social Interaction-based Features Interaction scenario. In: Proceedings of the 2009 International Conference on Multimodal Interfaces. Cambridge, MA, USA; 2009. p. 119–125.

75. Bartneck C, Kanda T, Mubin O, Al Mahmud A. Does the design of a robot influence its animacy and perceived intelligence? International Journal of Social Robotics. 2009; 1(2):195–204. https://doi.org/10.1007/s12369-009-0013-7

76. Richman WL, Kiesler S, Weisband S, Drasgow F. A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. Journal of Applied Psychology. 1999; 84:754–775. https://doi.org/10.1037/0021-9010.84.5.754

77. Baxter P, Kennedy J, Senft E, Lemaignan S, Belpaeme T. From Characterising Three Years of HRI to Methodology and Reporting Recommendations. In: HRI 2016. Christchurch, New Zealand: ACM Press; 2016. p. 391–398.

78. U.K. Department for Education School Performance Tables; 2015. http://www.education.gov.uk/schools/performance/

79. Baxter P, Ashurst E, Kennedy J, Senft E, Lemaignan S, Belpaeme T. The Wider Supportive Role of Social Robots in the Classroom for Teachers. In: 1st Int. Workshop on Educational Robotics at the Int. Conf. Social Robotics. Paris, France; 2015.

80. Dillenbourg P. Design for Classroom Orchestration. Computers and Education. 2013; 69:485–492. https://doi.org/10.1016/j.compedu.2013.04.013

frontiers in ICT

ORIGINAL RESEARCH
published: 24 April 2017
doi: 10.3389/fict.2017.00006

The Impact of Robot Tutor Nonverbal Social Behavior on Child Learning

James Kennedy[1]*, *Paul Baxter*[2] *and Tony Belpaeme*[1,3]

[1] Centre for Robotics and Neural Systems, Faculty of Science and Engineering, Plymouth University, Plymouth, UK, [2] Lincoln Centre for Autonomous Systems, School of Computer Science, University of Lincoln, Lincoln, UK, [3] ID Lab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

Several studies have indicated that interacting with social robots in educational contexts may lead to a greater learning than interactions with computers or virtual agents. As such, an increasing amount of social human–robot interaction research is being conducted in the learning domain, particularly with children. However, it is unclear precisely what social behavior a robot should employ in such interactions. Inspiration can be taken from human–human studies; this often leads to an assumption that the more social behavior an agent utilizes, the better the learning outcome will be. We apply a nonverbal behavior metric to a series of studies in which children are taught how to identify prime numbers by a robot with various behavioral manipulations. We find a trend, which generally agrees with the pedagogy literature, but also that overt nonverbal behavior does not account for all learning differences. We discuss the impact of novelty, child expectations, and responses to social cues to further the understanding of the relationship between robot social behavior and learning. We suggest that the combination of nonverbal behavior and social cue congruency is necessary to facilitate learning.

Keywords: human–robot interaction, robot tutors, social behavior, child learning, nonverbal immediacy

OPEN ACCESS

Edited by:
Hatice Gunes,
University of Cambridge, UK

Reviewed by:
Shin'Ichi Konomi,
University of Tokyo, Japan
Khiet Phuong Truong,
University of Twente, Netherlands

***Correspondence:**
James Kennedy
james.kennedy@plymouth.ac.uk

Specialty section:
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in ICT

Received: *31 December 2016*
Accepted: *29 March 2017*
Published: *24 April 2017*

Citation:
Kennedy J, Baxter P and Belpaeme T (2017) The Impact of Robot Tutor Nonverbal Social Behavior on Child Learning. Front. ICT 4:6. doi: 10.3389/fict.2017.00006

1. INTRODUCTION

The efficacy of robots in educational contexts has been demonstrated by several researchers when compared to not having a robot at all and when compared to other types of media, such as virtual characters (Han et al., 2005; Leyzberg et al., 2012; Tanaka and Matsuzoe, 2012; Alemi et al., 2014). One suggestion for why such differences are observed stems from the idea that humans see computers as social agents (Reeves and Nass, 1996) and that robots have increased social presence over other media as they are physically present in the world (Jung and Lee, 2004; Wainer et al., 2007). If the social behavior of an agent can be improved, then the social presence will increase and interaction outcomes should improve further (for example, through social facilitation effects (Zajonc, 1965)), but it is unclear how robot social behavior should be implemented to achieve such aims.

This has resulted in researchers exploring various aspects of robot social behavior and attempting to measure the outcomes of interactions in educational contexts, but a complex picture is emerging. While plenty of literature is available from pedagogical fields which describe teaching concepts, there are rarely examples of guidance for social behavior at the resolution required by social roboticists for designing robot behavior. The importance of social behavior in teaching and learning has been demonstrated between humans (Goldin-Meadow et al., 1992, 2001), but not enough is known for implementation in human–robot interaction (HRI) scenarios. This has led researchers to start exploring precisely how a robot should behave socially when information needs to be communicated

to, and retained by, human learners (Huang and Mutlu, 2013; Kennedy et al., 2015d).

In this article, we seek to establish what constitutes appropriate social behavior for a robot with the aim of maximizing learning in educational interactions, as well as how such social behavior might be characterized across varied contexts. First, we review work conducted in the field of HRI between robots and children in learning environments, finding that the results are somewhat mixed and that it is difficult to draw comparisons between studies (Section 2.1). Following this, we consider how social behavior could be characterized, allowing for a better comparison between studies and highlighting *immediacy* as one potentially useful metric (Section 2). Immediacy literature is then used to generate a hypothesis for educational interactions between robots and children. In an evaluation to test this hypothesis, nonverbal immediacy scores are gathered for a variety of robot behaviors from the same context (Section 3). While the data broadly agrees with the predictions from the literature, there are important differences that are left unaccounted for. We discuss these differences and draw on the literature to hypothesize a possible model for the relationship between robot social cues and child learning (Section 2.5). The work contributes to the field by furthering our understanding of the impact of robot nonverbal social behavior on task outcomes, such as learning, and by proposing a model that generates predictions that can be objectively assessed through further empirical investigation.

2. RELATED WORK

2.1. Robot Social Behavior and Child Learning in HRI

There are many examples of compelling results, which support the notion that the physical presence of a robot can have a positive impact on task performance and learning. Leyzberg et al. (2012) found that adults who were tutored by a physical robot significantly outperformed those who interacted with a virtual character when completing a logic puzzle. A controlled classroom-based study by Alemi et al. (2014) employed a robot to support learning English from a standard textbook over 5 weeks with a (human) teacher. In one condition, normal delivery was provided, and in the other, this delivery was augmented with a robot that was preprogrammed to explain words through speech and actions. It was found that using a robot to supplement teaching over this period led to significant child learning increases when compared to the same material being covered by the human teacher without a robot. This is strong evidence for the positive impact that robots can have in education, which has been supported in other scenarios. Tanaka and Matsuzoe (2012) also found that children learn significantly more when a robot is added to traditional teaching, both immediately after the experiment and after a delayed period (3–5 weeks later). Combined, these findings suggest that the use of a physically embodied robot can positively contribute to child learning.

Aspects of a robot's nonverbal social behavior have been investigated in one-on-one tutoring scenarios with mixed results. Two studies in the same context by Kennedy et al. (2015c) and Kennedy et al. (2015d) have found that the nonverbal behavior of a robot does have an impact on learning, but that the effect is not always in agreement with predictions from the human–human interaction (HHI) literature. These studies will be considered in more detail in Section 3. Similarly, Herberg et al. (2015) found that the HHI literature would predict an increase in learning performance with increased gaze of a robot toward a pupil, but the opposite was observed: an Aldebaran NAO would look either toward or away from a child while they completed a worksheet based on material they had learnt from the robot, but this was not found to be the case. However, Saerbeck et al. (2010) varied socially supportive behaviors of a robot in a novel second language learning scenario. These behaviors included gestures, verbal utterances, and emotional expressions. Children learnt significantly more when the robot displayed these socially supportive behaviors.

The impact on child learning of verbal aspects of robot behavior has also been investigated. Gordon et al. (2015) developed robot behaviors to promote curiosity in children with the ultimate aim of increased learning. While the children were reciprocal in their curiosity, their learning did not increase as the HHI literature would predict. Kanda et al. (2012) compared a "social" robot to a "non-social" robot, operationalized through verbal utterances to children when they are completing a task. Children showed a preference for the social robot, but no learning differences were found.

Ultimately, it is a difficult task to present a coherent overview of the effect of robot social behavior on child learning, with many results appearing to contradict one another or not being comparable due to the difference in learning task or behavioral context. More researchers are now using the same robotic platforms and peripheral hardware than before (quite commonly the Aldebaran NAO with a large touchscreen, e.g., Baxter et al. (2012)), but there remain few other similarities between studies. Behavior of various elements of the system is reported alongside learning outcomes, but it is difficult to translate from these descriptions to something that can be compared between studies. As such, it becomes almost impossible to determine if differing results between studies (and discrepancies with HHI predictions) are due to differences in robot behavior, the study population, other contextual factors, or indeed a combination of all three. It is apparent that a characterization of the robot social behavior would help to clarify the differences between studies and provide a means by which certain factors could be accounted for in analysis; this will be explored in the following section.

2.2. Characterizing Social Behavior through Nonverbal Immediacy

To allow researchers to make clearer comparisons between studies and across contexts, a metric to characterize the social behavior of a robot is desirable. Various metrics have been used before in HRI. Retrospective video coding has been used in several HRI studies as a means of measuring differences in human behavioral responses to robots, for example, the studies by Tanaka and Matsuzoe (2012); Moshkina et al. (2014); Kennedy et al. (2015b). However, this method of characterizing social behavior is incredibly time consuming, particularly when the coding of multiple social cues is required. Furthermore, it provides data

for social cues in isolation and does not easily provide a holistic characterization of the behavior. It is unclear what it means if the robot gazes for a certain number of seconds at the child in the interaction and also performs a certain number of gestures; this problem is exacerbated when a task context changes. The perception of the human directly interacting with the robot is also not accounted for. It is suggested that the direct perception of the human within the interaction is an important one, as they are the one being influenced by the robot behavior *in the moment*. This cannot be captured through *post hoc* video coding.

The Godspeed questionnaire series developed by Bartneck et al. (2009b) has been used in many HRI studies to measure users' perception of robots (Bartneck et al., 2009a; Ham et al., 2011). The animacy and anthropomorphism elements of the scale in particular consider the social behavior and perception of the robot. However, it is not particularly suited to use with children due to the language level (i.e., use of words such as "stagnant," "organic," and "apathetic"). It may also be that the questionnaire would measure aspects of the robot not directly related to social behavior as it is asking about more general perceptions. While this could be of use in many studies, for the aim of characterizing social behavior in the case here, these aspects prevent suitable application.

Nonverbal immediacy (NVI) was introduced in the 1960s by Mehrabian (1968) and is defined as the "psychological availability" of an interaction partner. Immediacy is further introduced as being a measure that indicates "the attitude of a communicator toward his addressee" and in a general form "the extent to which communication behaviors enhance closeness to and nonverbal interaction with another" (Mehrabian, 1968). A number of specific social behaviors are listed (touching, distance, forward lean, eye contact, and body orientation) to form a part of this measure, which were later utilized by researchers that sought to create and validate measuring instruments for NVI. However, it is also this feature that makes NVI a particularly enticing prospect for designers of robot behavior, as the social cues used in the measure are explicit (which is often not the case in other measures of perception commonly used in the field, e.g., Bartneck et al. (2009b)). A reasonable volume of data also already exists for studies considering immediacy, with over 80 studies (and N nearly 25,000) from its inception to 2001 (Witt et al., 2004) and more since. This provides a context for NVI findings in HRI scenarios and a firm grounding in the human–human literature from which roboticists can draw.

Several versions of surveys have been developed and validated for measuring the nonverbal immediacy of adults (Richmond et al., 2003). Surveys have also been developed for verbal immediacy (Gorham, 1988), but their ability to measure precisely the concept of verbal immediacy remains the subject of debate (Robinson and Richmond, 1995). Both verbal and nonverbal measures consider observed overt behavior more than, but not excluding, perceptions. Immediacy has recently been used in HRI as a means of motivating robot behavior manipulations (Szafir and Mutlu, 2012) and characterizing social behavior (Kennedy et al., 2017).

There is a consensus on the instruments used to measure nonverbal immediacy (whereas this is less clear for verbal immediacy), and it is also transparent in terms of how participants are judging the robot. The Godspeed questionnaire is a useful tool for gathering perceptions, but nonverbal immediacy is clearly measuring overt social behavior, and so it is ideal given our scope of trying to characterize social behavior (often with children). Use of the NVI metric brings several other advantages to researchers in HRI and for robot behavior designers. The NVI metric can be used as a guideline for an explicit list of social cues available for manipulation as a part of robot behavior. Characterization of robot social behavior at this relatively low level is not readily available in other metrics. This provides a useful first step in designing robot behavior but also a means of evaluating and modifying future social behaviors. NVI constitutes part of an overall social behavior; hence NVI is treated as a *characterization* of the overall behavior, not a complete description or definition. Not all aspects of sociality or interaction are addressed through the measure, but to the knowledge of the authors, nor are these aspects fully covered by any other validated metric.

The NVI metric can be used with either the subjects themselves or with observers (during or after the interaction). This permits flexibility depending on the needs of the researcher. It is not always practical to collect such data from participants (for example, when they are young children or following an already lengthy interaction), so having the flexibility to gather these data *post hoc* is advantageous. Due to this mixture of practical and theoretical benefits, nonverbal immediacy (NVI) will be adopted as a social behavior characterization metric for this article.

Immediacy has been validated through physical manipulation of some of the social cues, specifically eye gaze and proximity, to ensure that the phenomenon indeed works in practice and is not a product of affect or bias in survey responses (Kelley and Gorham, 1988). It was indeed found that the physical manipulations that were made which would lead to a higher immediacy score (standing closer and providing more eye gaze) did lead to increased short-term recall of information. While there is clearly a difference between recall and learning, recall of information is a promising first step to acquiring new understanding and skills. These results were hypothesized to exist in the other immediacy behaviors (such as gestures) as well. Overall, the link between teacher immediacy and student learning is hypothesized to be a positive one, as reflected in the meta-review by Witt et al. (2004) and many studies (Comstock et al., 1995; McCroskey et al., 1996; Christensen and Menzel, 1998). Thus, this prediction can be tested in human–robot interaction, where the robot takes the role of the tutor. As a result, we generate the following hypothesis:

H1. A robot tutor perceived to have higher immediacy leads to greater learning than a robot perceived to have lower immediacy.

3. APPLYING NONVERBAL IMMEDIACY TO HRI

In this section, an evaluation of nonverbal immediacy (NVI) in the context of cHRI is described. The aim is to explore whether the characterization that it provides can account for the differences between robot behaviors and learning outcomes of children. The wealth of literature that explores NVI in educational scenarios is

generally in agreement that higher NVI of an instructor is positively correlated with learning outcomes of students. We evaluate 4 differently motivated robot behaviors and a human in a one-to-one maths-based educational interaction with children. The aim is to use these data to provide a comparison between behavioral manipulations to test predictions from the HHI immediacy literature regarding social behavior.

3.1. Task Design and Measures

All five behaviors under consideration use the same context and broader methodology. Children aged 8–9 years are taught how to identify prime numbers between 10 and 100 using a variation on the Sieve of Eratosthenes method. They interact with a tutor: in 4 conditions, this is an Aldebaran NAO robot, and in 1 condition, this is a human (**Figure 1**). Children complete pretests and posttests in prime number identification, as well as pretests and posttests for division by 2, 3, 5, and 7 (skills required by the Sieve of Eratosthenes method for numbers in the range used) on a large touchscreen. The tutor provides lessons on primes and dividing by 2, 3, 5, and 7 (**Figure 2**). In all cases, an experimenter briefs the child and introduces the child to the tutor. The experimenter remains in the room throughout the interaction, but out of view of the child. Two cameras record the interactions; one is directed toward the child and one toward the tutor. Interactions with the tutor would last for around 10–15 min, with an additional 5 min required afterward in conditions where nonverbal immediacy surveys were completed (details to follow).

At the start of the interaction, the children complete a pretest in prime numbers on the touchscreen without any feedback from the screen or the tutor. A posttest is completed by the children at the end of the interaction; again no feedback is provided to the child so as not to influence their categorizations. Two tests are used in a cross-testing strategy, so children have a different pretest and posttest, and the tests are varied as to whether they are used as a pretest or posttest. The tests require the children to categorize numbers as "prime" or "not prime" by dragging and dropping numbers on screen into the category labels. Each test has 12 numbers, so by chance, a score of 6 would be expected (given 2 possible categories 50% is chance). Learning is measured through the improvement in child score from the prime number pretest to posttest. By considering the improvement, any prior knowledge (correct or otherwise) or deviation in division skill is factored in to the learning measure. The mean and *SD* score (of 12) for the pretests are compared to those of the posttest to calculate the learning effect size (Cohen's *d*) for each condition.

The prime number task was selected in consultation with education professionals to ensure that it was appropriate for the capabilities of children of this age. Children of this age have not yet learnt prime number concepts in school, but do have sufficient (but imperfect) skills for dividing by 2, 3, 5, and 7 as required by the technique for calculating whether numbers are prime. During the division sections of the interaction, the tutor provides feedback on child categorizations.

Nonverbal immediacy (NVI) scores are collected through questionnaires. For children, this was done after the interaction with the tutor had been completed, for adults, this was online (details in Section 3.4). A standard nonverbal immediacy questionnaire was adapted for use with children by modifying some of the language; the original and modified versions alongside the score formula can be seen online.[1] Both the Robot Nonverbal Immediacy Questionnaire (RNIQ) and Child-Friendly Nonverbal Immediacy Questionnaire (CNIQ) were used depending on condition for children. Adults had the same questionnaire but with "the child" in place of "you" as they were observing the interaction, rather than participating in it. The questionnaire consists of 16 questions about overt nonverbal behavior of the tutor. Each question is answered on a 5-point Likert scale, and a final immediacy score is calculated by combining these answers. Some count positively toward the nonverbal immediacy score, whereas some count negatively, depending on the wording of the question. The version in the Appendix shows the questionnaire used for this study when a robot (as opposed to a human) tutor was used as this has been validated for use in HRI (Kennedy et al., 2017) and corresponds to the validated version from prior human-based literature (Witt et al., 2004).

Existing immediacy literature extensively uses adults (often students) as subjects; studies with children are rare. Prior work

[1] http://goo.gl/UoL5QM, also included as an Appendix.

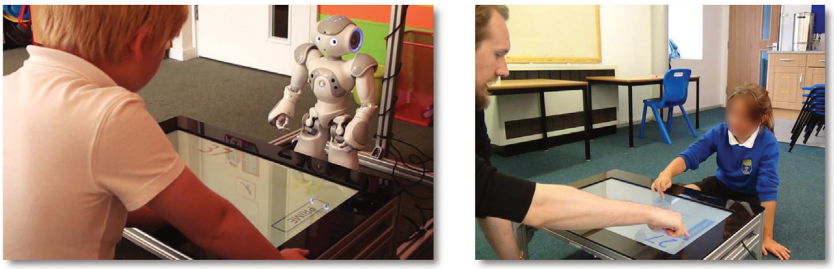


FIGURE 1 | (Left) Still image from a human–robot interaction (specifically, the "social" condition), and (right) still image from the human–human condition. The tutor (either robot or human) teaches children how to identify prime numbers using the Sieve of Eratosthenes method using a large horizontal touchscreen as a shared workspace. The robot can "virtually" move numbers on screen (numbers move in correspondence with robot arm movements, but physical contact is not made with the screen).

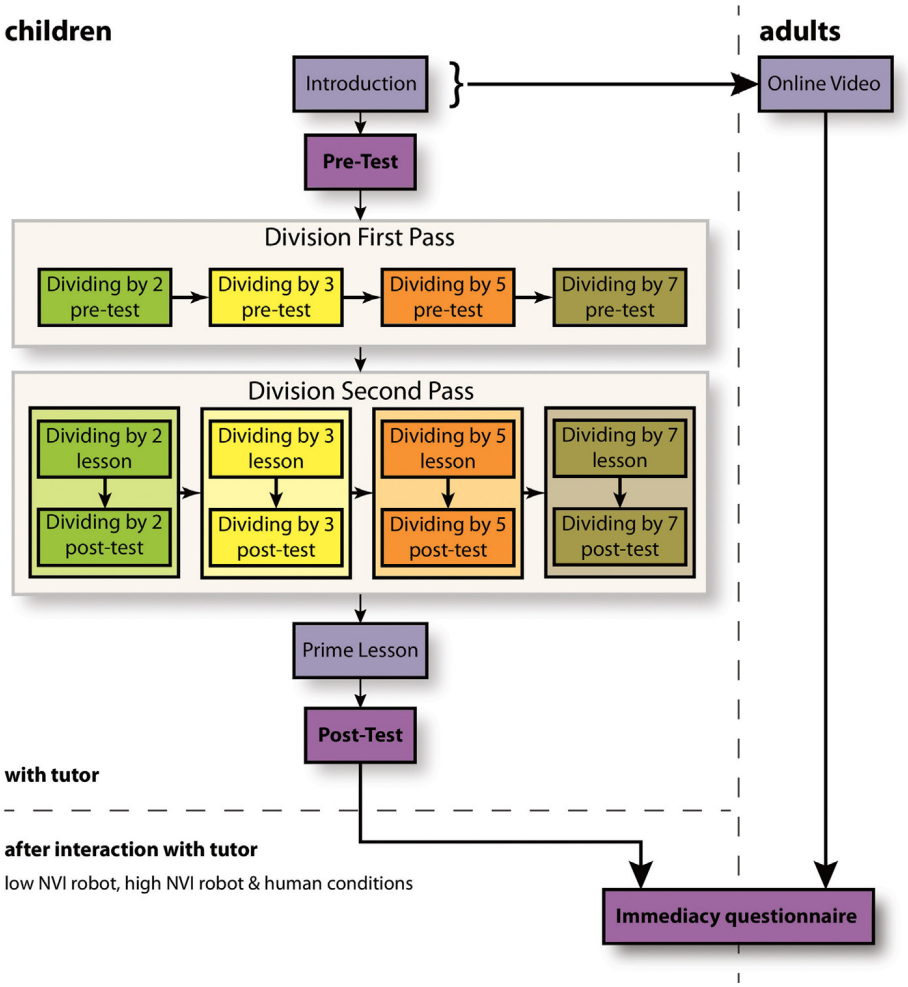


FIGURE 2 | Task structure—the top section is led by the tutor and is aimed at teaching children how to calculate whether a number is prime. The bottom section consists of completing the nonverbal immediacy questionnaire—this is done after the interaction for 3 of the child conditions and *via* online videos to get adult responses. Dark purple boxes (pretest, posttest, and immediacy questionnaire) are the metrics under consideration in this article.

has been conducted with the adapted nonverbal immediacy scale for use with robots and children (Kennedy et al., 2017); however, the task in this article is novel in this context (one-to-one interactions instead of group instruction). Children present unique challenges when using questionnaire scales, such as providing different answers for negatively worded questions to positively worded ones (Borgers et al., 2004) or trying to please experimenters (Belpaeme et al., 2013), which can consequently make it difficult to detect differences in responses (Kennedy et al., 2017). As children are not well represented in immediacy literature, using adults for NVI scores more tightly grounds our hypotheses and assumptions to the existing literature. However, NVI ratings are collected from children in robot conditions in which NVI is intentionally manipulated. As the nonverbal immediacy was intentionally manipulated between these conditions, and the adult results can provide some context, we can observe whether children do perceive the manipulation on this scale, potentially broadening the applicability of our findings.

3.2. Conditions

A total of 5 conditions are used in this evaluation.[2] As described in the introduction, an often adopted approach to social behavioral design is to consider how a human behaves and reproduce that (insofar as is possible) on the robot. As such, we use 2 conditions, seeking to follow and also invert this approach. We additionally use 2 conditions derived from the NVI literature, again seeking to maximize and minimize the behaviors along this scale. The final condition is a human benchmark. Further details for each can be seen in **Table 1** and below:

1. "Social" robot (SR)—this condition is derived from observations of an expert human–human tutor completing this task with 6 different children. This condition reflects a human

[2]Please note that while some data have previously been published for all of these conditions (Kennedy et al., 2015c,d, 2016), this article presents both novel data collection and different analysis perspectives in a new context to the prior work.

TABLE 1 | Operationalization of the differences in nonverbal behavior between the conditions considered in the study presented in this article.

Condition	Motivation	Nonverbal behavior	Other manipulations
"Social" robot (SR)	Based on a human model of the task	Seeks mutual gaze with child, frequent arm gestures	Uses child name, personalizes number of items in division posttests, "positive" feedback, variable feedback
"Asocial" robot (AR)	"Inverse" of the above human model	Avoids child gaze, frequent but mistimed arm gestures	Blunt feedback, repetitive feedback
High NVI robot (HNVI)	Intended to maximize the nonverbal immediacy	Seeks mutual gaze with child, frequent head/gaze movement, frequent arm gestures, lean forwards, continuous small upper body movements	
Low NVI robot (LNVI)	Intended to minimize the nonverbal immediacy	Avoids child gaze, infrequent head/gaze movement, no arm gestures, TTS parameters modified to give "dull" voice, lean backward, rigid/no upper body movements	
Human (HU)	Human benchmark	No instructions given for nonverbal behavior	

Further notes are provided about any other manipulations made besides nonverbal behavior.

model-based approach to designing the behavior. The social behavior of the tutor was analyzed through video coding, and these behaviors were implemented on the robot where possible.

2. "Asocial" robot (AR)—this condition considers the behavior generated for the SR condition and seeks to "invert" it. That is, the behavior is intentionally manipulated such that an opposite implementation is produced, for example, the SR condition seeks to maximize mutual gaze, whereas this condition actively minimizes mutual gaze. The quantity of social cues used in this condition is exactly the same as the SR condition above; however, the placement of these cues is varied (for example, a wave would occur during the greeting in SR, but during an explanation in AR).
3. High NVI robot (HNVI)—this condition uses the literature to drive the behavioral design. The behavior is derived from considering how the social cues within the nonverbal immediacy scale can be maximized. For example, the robot will seek to maximize gaze toward the child and make frequent gestures.
4. Low NVI robot (LNVI)—this condition is intended to be the opposite to the HNVI condition. Again, the nonverbal immediacy literature is used to drive the design, but in this case, all of the social cues are minimized. For example, the robot avoids gazing at the child and makes no gestures.
5. Human (HU)—this is a human benchmark. The human follows the same script for the lessons as the robot, but they are not constrained in their social behavior. The intention here is that we can then acquire data for a "natural", non-robot interaction where the social behavior is not being manipulated; this can then be used to provide context for the robot conditions.

A summary of the motivations for the conditions and the operationalization of the differences between conditions can be seen in **Table 1**. Further implementation details can be seen in "Robot Behavior." While the Aldebaran NAO platform cannot be manipulated for some of the cues involved in the nonverbal immediacy measure given the physical setup and modalities of the robot (i.e., smiling and touching), it has been manipulated on all of the other cues possible. This leaves only 4 of the 16 questions (2 of 8 cues) not manipulated in the metric. Specifically, these are questions 4, 8, 9, and 13, as seen in the Appendix, pertaining to frowning/smiling and touching.

3.2.1. Robot Behavior

Throughout the division sections of the interaction, the tutor (human or robot) would provide feedback on child categorizations and could also suggest numbers for the child to look at next. This was done through moving a number to the center of the screen and making a comment such as "why don't you try this one next?" The tutor would also provide some prescripted lessons (**Figure 2**) that would include 2 example categorizations on screen. These aspects are central to the delivery of the learning content, so are maintained across all conditions to prevent a confound in learning content.

All robot behavior was autonomous, apart from the experimenter clicking a button to start the system once the child was sat in front of the touchscreen. The touchscreen and a Microsoft Kinect were used to provide input for the robot to act in an autonomous manner. The touchscreen would provide information to the robot about the images being displayed and the child moves on screen, the Kinect would provide the vector of head gaze for the child and whether this was toward the robot. Through these inputs, the robot behavior could be made contingent on child actions, for example, by providing verbal feedback after child moves (in all conditions), or manipulating mutual gaze. In all robot conditions, the robot gaze was contingent on the child's gaze, but with differing strategies depending on the motivation of the condition. The AR and LNVI conditions would actively minimize mutual gaze by intentionally avoiding looking at the child, whereas the SR and HNVI conditions would actively maximize mutual gaze by looking at the child when data from the Kinect indicated that the child was looking at the robot. Robot speech manipulation executed in the LNVI condition to make the robot voice "dull" was achieved through lowering the vocal shaping parameter of the TTS engine (provided by Acapela).

Due to the human model-based approach, some personalization aspects such as use of child name were included as part of the social behavior in the SR condition. This was not done in the NVI conditions as these manipulations are not motivated through the NVI metric. The HNVI condition also addresses more of the NVI questionnaire items (leaning forward and continuous "relaxed" upper body movements) than the SR condition due to this difference in motivation. The AR condition has the same quantity

of behavior as the SR condition, whereas the LNVI has a *lack* of behavior. As a concrete example, the AR condition includes inappropriately placed gestures, whereas the LNVI condition includes no gestures. Consequently, the LNVI and HNVI conditions provide useful comparisons both to one another and to the SR and AR conditions.

3.3. Participants

To provide NVI scores for all 5 conditions, video clips of the conditions were rated by adults. Nonverbal immediacy scores were also acquired at the time of running the experiments for 3 of the 5 conditions (high and low NVI robot and human) from children through paper questionnaires (**Table 2**). These scores allow a check that the NVI manipulation between the robot conditions could be perceived by the children, with the adult data provided context for these ratings. Written informed consent from parents/guardians was received for the children to take part in the study, and they additionally provided verbal assent themselves, in accordance with the Declaration of Helsinki. Written informed consent from parents/guardians and verbal assent from children were also received for the publication of identifiable images. The protocol was reviewed and approved by the Plymouth University ethics board. **Table 2** shows numbers of participants per condition and average ages for the adult conditions; all children were aged 8 or 9 years old and were recruited through a visit to their school, where the experiment took place.

3.4. Adult Nonverbal Immediacy Score Procedure

Videos shown to adults to acquire nonverbal immediacy scores were each 47 s long. The videos contained both the interaction video (42 s) and a verification code (5 s; details in the following paragraph). The length of video was selected to be 42 s as the literature suggests that at least around 6 s are required to form a judgment of social behavior (Ambady and Rosenthal, 1993), and there was a natural pause at 42 s in the speech in all conditions so that it would not cut part-way through a sentence. The interaction clips were all from the start of an interaction, so the same information was being provided by the tutor to the child in the clip.

To provide sufficient subject numbers for all of the conditions, an online crowdsourcing service[3] was used. The participants were restricted to the USA and could only take part if they had a reliable record within the crowdsourcing platform. A test question was put in place whereby participants had to enter a 4 digit number into a text box. This number was shown at the end of the video for 5 s (the video controls were disabled so it could not be paused and the number would disappear after the video had finished). A different number was used for each video. If the participants did not enter this number correctly, then their response was discarded. The crowdsourcing platform did not allow the prevention of users completing multiple conditions, so any duplicates were removed, i.e., only those seeing a video for the first time were kept as valid responses. A total of 366 responses were collected, but 209 were discarded as they did not answer the test question correctly, the user had completed another condition,[4] or the response was clearly spam (for example, all answers were "1"). This left 157 responses across 5 conditions; 90M/67F (**Table 2**).

4. RESULTS

When performing a one-way ANOVA, a significant effect is found for condition seen, showing that the robot behavior influences perceived nonverbal immediacy; $F(4,152) = 14.057$, $p < 0.001$. *Post hoc* pairwise comparisons with Bonferroni correction reveal that the adult-judged NVI of the LNVI condition is significantly different to all other conditions ($p < 0.001$ in all cases), but no other pairwise comparisons are statistically significant at $p < 0.05$. The nonverbal immediacy score means and learning effect sizes for each condition can be seen in **Table 3**. Children learning occurs in all conditions. Generally, it can be seen that the conditions with higher rated nonverbal immediacy lead to greater child improvement in identifying prime numbers.

While significance testing provides an indication that most of the conditions are similar (at least statistically) in terms of NVI, additional information for addressing the hypothesis can be gleaned by considering the trend that these data suggest (**Figure 3**). A strong positive correlation is found between the (adult) NVI score of the conditions and the learning effect sizes (Cohen's *d*) of children who interacted in those conditions ($r(3) = 0.70$, $p = 0.188$). This correlation is not significant, likely due to the small number of conditions under consideration, but the strength of the correlation suggests that a relationship could be present.

[3] http://www.crowdflower.com/.

[4] The majority of exclusions were due to users having completed another condition, thereby impairing the independence of the results.

TABLE 2 | Subject numbers by condition and average ages for adult participants by condition.

Condition	Child *N*	Adult *N*	Adult *M* age, *SD* in brackets	Child immediacy scores collected?
Low NVI robot	12	33	31.5 (*12.2*)	Yes
High NVI robot	11	31	35.6 (*11.7*)	Yes
Social robot	12	33	29.0 (*10.4*)	No
Asocial robot	11	30	39.0 (*12.2*)	No
Human	11	30	32.9 (*12.3*)	Yes

TABLE 3 | Adult and child nonverbal immediacy ratings and child learning (as measured through effect size between pretests and posttests for prime numbers) by tutor condition.

Condition	Adult *M* NVI rating *[95% CI]*	Child *M* NVI rating *[95% CI]*	Child learning (*d*)
Low NVI robot	40.2 *[38.1, 42.2]*	51.0 *[47.6, 54.4]*	0.30
High NVI robot	48.4 *[46.9, 50.0]*	55.1 *[52.3, 57.6]*	0.67
Social robot	49.0 *[47.6, 50.4]*	N/A	0.51
Asocial robot	48.5 *[46.1, 50.8]*	N/A	0.89
Human	47.7 *[45.3, 50.1]*	54.4 *[52.9, 55.9]*	0.89

When the immediacy scores provided by the children who interacted with the robot are also considered, a similar pattern can be seen (**Figure 4**). The adult and child immediacy ratings correlate well, with a strong positive correlation ($r(1) = 1.00$, $p < 0.001$). There is also a strong positive correlation for the children between immediacy score of the conditions and the learning effect sizes (Cohen's d) in those conditions ($r(1) = 0.86$, $p = 0.341$). Again, significance is not observed, but the power of the test is low due to the number of data points available for comparison. The strong positive correlations between child immediacy scores and learning and adult immediacy scores and learning provide some support for hypothesis H1 (that higher tutor NVI leads to greater learning), but further data points would be desired to explore this relationship further. It should be noted that we consider the results of 57 children and 157 adults across 5 conditions; acquiring further data points for more behaviors (and deciding what these behaviors should be) would be a time-consuming task.

5. DISCUSSION

There is a clear trend in support of hypothesis H1: that a tutor perceived to have higher immediacy leads to greater learning. As such, increasing the nonverbal immediacy behaviors used by a social robot would likely be an effective way of improving child learning in educational interactions. However, nonverbal immediacy does not account for all of the differences in learning. Three of the conditions have near identical NVI scores as judged by adults, but quite varied learning results (high NVI robot: $M = 48.4$ NVI score/$d = 0.67$ pre–post test improvement, asocial robot: NVI $M = 48.5$/$d = 0.89$, social robot: NVI $M = 49.0$/$d = 0.51$). This partially reflects the slightly mixed

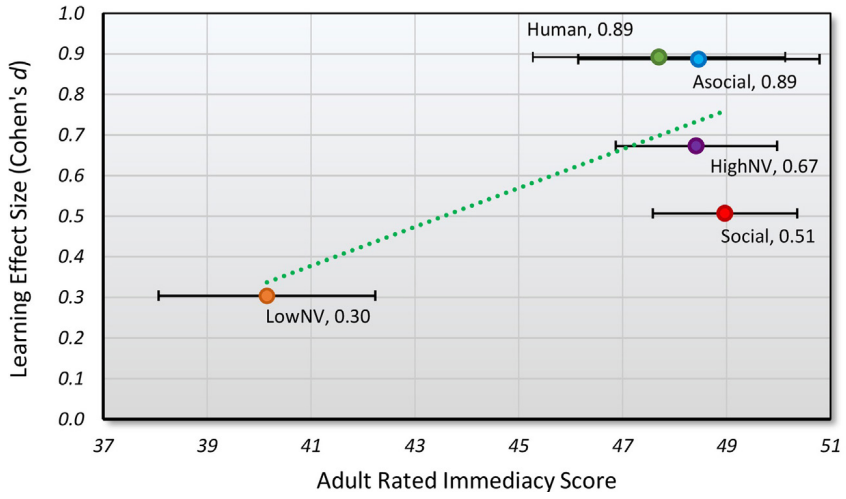


FIGURE 3 | **Nonverbal immediacy scores as judged by adults and learning effect sizes for the prime number task**. The dotted green line indicates a trend toward greater nonverbal immediacy of the tutor leading to increased learning. Error bars show 95% confidence interval.

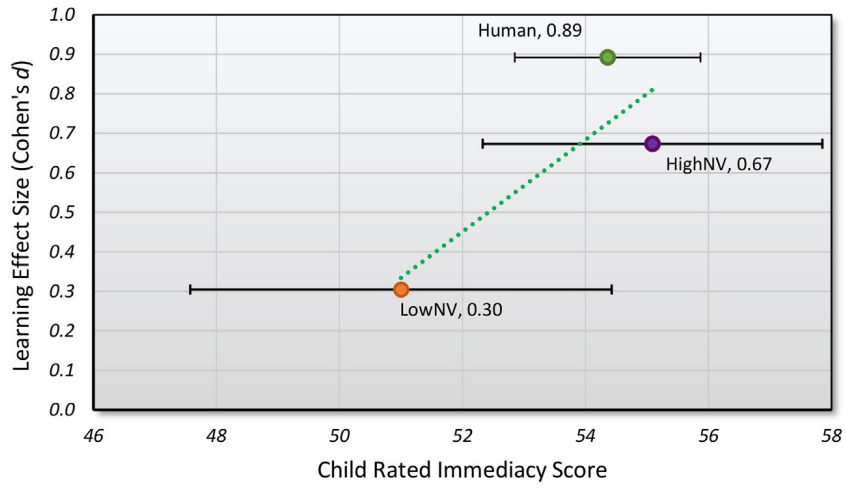


FIGURE 4 | **Nonverbal immediacy scores as judged by the children in the interaction and learning effect sizes for the prime number task**. The dotted green line indicates a trend toward greater perceived nonverbal immediacy of the tutor leading to increased learning. Error bars show 95% confidence interval.

picture of immediacy that the pedagogy literature presents; for example, the disagreement as to whether NVI has a linear (Christensen and Menzel, 1998) or curvilinear (Comstock et al., 1995) relationship with learning. Nonetheless, there are further factors that may be introduced by the use of a robot that may have had an influence on the results. Nonverbal immediacy only considers overt observed social behaviors, so by design does not cover all possible aspects of effective social behavior for teaching. While this seems to be enough in HHI (Witt et al., 2004), it may not be for HRI since various inherent facets of human behavior cannot be assumed for robots. Several possible explanations as to why this learning variation is present will now be discussed. From this, a possible model (suggested to be more accurate) of the relationship between social behavior and learning is proposed. Such a model may be useful in describing (and testing) the relationship between social behavior and child learning for future research.

5.1. Timing of Social Cues

The quantity of social cues used in both the social robot and the asocial robot conditions is exactly the same; however, the timing is varied. Timing is not considered as part of the nonverbal immediacy metric—the scale measures whether cues have, or have not, been used, rather than whether their timing was appropriate. The cues used in the asocial robot condition were intentionally placed at inappropriate times (for example, waving part-way through the introduction, instead of when saying hello). This is not factored into the nonverbal immediacy measure, but could impact the learning (Nussbaum, 1992).

The timing of social cues in the human condition may also explain why the learning in this condition was higher than the others. The robot conditions are contingent on aspects of child behavior, such as gaze and touchscreen moves, but are not adapted to individual children (for example, the number of feedback instances the robot provides would not be based on how well the child was performing). However, the human is presumably adaptive in both the number of social cues used and the timing of these cues. Again, this would not be directly revealed by the immediacy metric, but could account for some of the learning difference. Indeed, the nonverbal immediacy metric comes from HHI studies and has been validated in such environments. In HHI, there is a reasonable assumption that the timing of social cues will be appropriate, and so it may not be necessary to include it as part of a behavioral metric for HHI. However, when applied to social robotics, the assumption of appropriate timing no longer applies, and so to fully account for learning differences in HRI, timing may need more explicit incorporation into characterizations of social behavior. This constitutes a limitation of the NVI metric, but also an opportunity for expansion in future work to capture timing aspects.

5.2. Relative Importance of Social Cues

One substantial difference between the robot conditions and the human condition is the possibility of using facial expressions. The robotic platform used for the studies was the Aldebaran NAO. This platform has limited ability to generate facial expressions as none of the elements of the face can move, only the eye color can be changed. On the other hand, the human has a rich set of facial expressions to draw upon.

While the overall nonverbal immediacy scores for the asocial, social, and human conditions are tightly bunched, the make-up of the scores is not. For example, the robot scores (asocial and social combined) are higher for gesturing, averaging $M = 4.3$ (95% CI 4.1, 4.5) out of 5 for the nonverbal immediacy question about gesturing (the robot uses its hands and arms to gesture while talking to you), compared to $M = 3.1$ (95% CI 2.7, 3.5) for the human. However, the human is perceived to smile more ($M = 2.5$, 95% CI 2.1, 2.8) than the robot ($M = 1.8$, 95% CI 1.5, 2.0). Through principle component analysis, Wilson and Locker (2007) found that different elements of nonverbal behavior do not contribute equally to either the nonverbal immediacy construct or instructor effectiveness. Facial expressions (specifically smiles) have a large impact on both the nonverbal immediacy construct and the instructor effectiveness, whereas gestures do not have such a large effect (although still a meaningful contribution; smiles: 0.54, gestures: 0.30 component contribution from Wilson and Locker (2007)).

In the nonverbal immediacy metric, all social cues are given equal weighting. However, this may not always be the most appropriate method for combining the cues given the evidence, which suggests that some cues may contribute more than others to various outcomes (McCroskey et al., 1996; Wilson and Locker, 2007). This could be a further explanation as to why several of the conditions in the study conducted here have near identical overall nonverbal immediacy scores, but very different learning outcomes.

5.3. Novelty of Character and Behavior

The novelty of both the character (i.e., robot or human) and the behavior itself could have had an impact on the learning results found in the study. Novelty is often highlighted as a potential issue in experiments conducted in the field (Kanda et al., 2004; Sung et al., 2009). The novelty of the robot behavior could override the differences between the conditions and subsequently influence the learning of the child. In the social robot condition here, novel behavior (such as new gestures) was often introduced when providing lessons to the child. Between humans, this would likely result in a positive effect (Goldin-Meadow et al., 2001), but when done by a robot, the novelty of the behavior may counteract the intended positive effect.

There may also be a difference in the novelty effect for the children seeing the robot when compared to the human. Although the human is not one that they are familiar with, they are still "just" a human, whereas the robot is likely to be more exciting and novel as child interaction with robots is more limited than with humans. The additional novelty of the robot could have been a distraction from the learning, explaining why the learning in the human condition is higher.

Finally, the novelty may have impacted the nonverbal immediacy scores themselves. It is possible that observers (be they children or adults) score immediacy on a relative scale. It is reasonable to suggest that the immediacy of the characters is judged not as a standalone piece of behavior, but in the context of an observer's prior experience, or expectations for what that character may be

capable of. Clear expectations will likely exist for human behavior, but not for robot behavior, which may lead to an overestimation of robot immediacy. This would impact on the ability of considering the human and robots on the same nonverbal immediacy scale and drawing correlations with learning and cannot be ruled out as a factor in the results.

5.4. (In)Congruency of Social Cues

As previously discussed, the robot is limited in the social cues that it can produce (for example, it cannot produce facial expressions). This meant that the conditions all manipulated the available robot social cues, but if social cues are interpreted as a single percept by the human (as suggested by the literature (Zaki, 2013)), then this could lead to complications.

In the case of the social robot, many social cues are used to try and maximize the "sociality" of the robot. This means that there is a lot of gaze from the robot to the child, and the robot uses a lot of gestures. However, it still cannot produce facial expressions. This incongruency between the social cues could produce an adverse effect in terms of perception on the part of the child and subsequently diminish the learning outcome. There are clear parallels here with the concept of the Uncanny Valley (Mori et al., 2012), with models for the Uncanny Valley based on category boundaries in perception indicating issues arising from these mismatches (Moore, 2012).

The expectation the child has for the robot social behavior is suggested to be of great importance (Kennedy et al., 2015a). If their expectations are formed early on through high quantities of gaze and gestures, then there would be a discrepancy when facial expressions do not match this expectation. Again, this expectation discrepancy may lead to adverse effects on learning outcomes, as in the case of perceptual issues due to cue incongruence. These issues may become exacerbated as the overall level of sociality of behavior of the robot increases as any incongruencies then become more pronounced. As stated in the study by Richmond et al. (1987), higher immediacy generally leads to more communication, which can create misperceptions (of liking, or expected behavior).

As the nonverbal immediacy scale has been rigorously validated (McCroskey et al., 1996; Richmond et al., 2003), it is known that it does indeed provide a reliable metric for immediacy in humans (Cronbach's alpha is typically between 0.70 and 0.85 (McCroskey et al., 1996)). Typically, internal consistency measures of a scale would be used to evaluate the ability of items in a scale to measure a unidimensional construct, i.e., how congruent the items are with one another. As such, a consistency measure could be used as an indicator of the congruency between the cues. The robot lacks a number of capabilities when compared to humans, and there are several scale items that are known to be impaired on the robot, such as smiling/frowning. Using an internal consistency measure across all NVI questionnaire items (with the negatively worded question responses reversed) can reveal cases in which the cues are relatively more or less congruent. Greater internal consistency indicates lower variability between questionnaire items (the social cues) and, therefore, more congruence between the social cues. Lower internal consistency indicates larger variability between scale items and thus greater incongruency between the cues.

Guttman's λ_6 (or G6) for each condition has been calculated,[5] revealing that indeed there are differences in how congruent the cues could be considered to be (**Table 4**; **Figure 5**). All of the NVI questionnaire items are included in the λ_6 calculation. The behavioral conditions used here are restricted in such a way that a lower reliability would be expected (as several cues of the scale are not utilized) for some conditions. Indeed, these values fall in line with predictions that could be made based on the social behavior in each of the conditions. The human reliability score provides a "sanity check" as it is assumed that human behavior would have a certain degree of internal consistency between social cues, which is reflected by it having the highest value. In addition, the LNVI robot condition has intentionally low NVI behavior, so the lack of smiling or touching (high NVI behaviors) does not cause incongruency (signified by a lower λ_6 score), whereas the HNVI robot condition has intentionally high NVI behavior where possible on the robot, so the lack of smiling and touching cause greater overall incongruency, resulting in a considerably lower λ_6 score.

5.5. A Hypothesis: Social Cue Congruency and Learning

Taking Guttman's λ_6 to provide an indication of the congruency of social cues, then it is clear that this alone would not provide a strong predictor of learning (**Figure 5**). However, these data can be combined with the social behavior (as measured through immediacy) to be compared to learning outcomes. In the resulting space, both congruency and social behavior could have an impact on learning, as hypothesized in the previous section (**Figure 6**).

Our data show that learning is best with human behavior, which is shown to be highly social and reasonably congruent. When the social behavior used is congruent, but not highly social, then the learning drops to a low level. The general trend of our data shows that when the congruency of the cues increases

[5]Cronbach's alpha tends to be the de facto standard for evaluating internal consistency and reliability; however, its use as such a measure has been called into question (Revelle and Zinbarg, 2009)—including by its own creator (Cronbach and Shavelson, 2004). Instead λ_6 is used, which considers the amount of variance in each item that can be accounted for by the linear regression of all other items (the squared multiple correlation) (Guttman, 1945). This provides a lower bound for item communality, becoming a better estimate with increased numbers of items. This would appear to provide a logical (but likely imperfect) indicator for the congruency of cues as required here.

TABLE 4 | Guttman's λ_6 and learning effect size by condition.

Condition	Learning effect size (Cohen's *d*)	Guttman's λ_6 (G6)
Asocial robot	0.89	0.84
Social robot	0.51	0.83
High NVI robot	0.67	0.69
Low NVI robot	0.30	0.78
Human	0.89	0.87

λ_6 is used as an indicator of social cue congruency, with a higher value indicating greater congruency between cues.

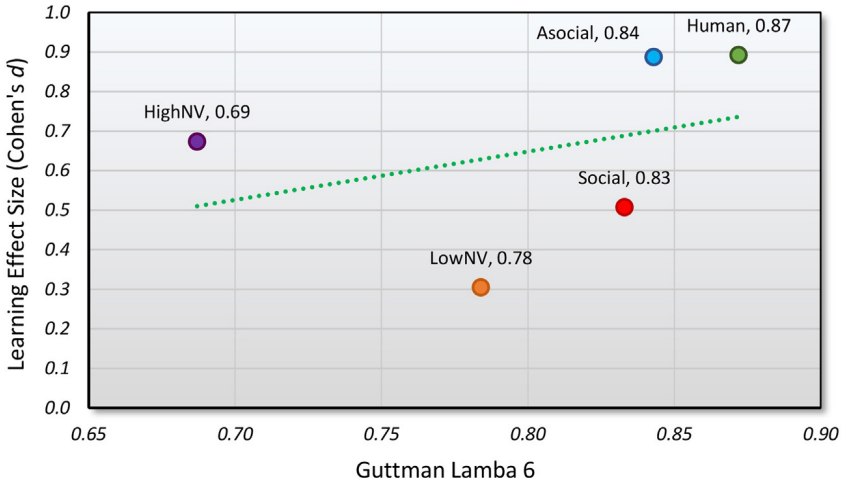


FIGURE 5 | Guttman's λ_6 against learning effect size for each of the prime tutoring conditions. The dotted line indicates a trend toward greater internal consistency (measured through λ_6) leading to greater learning.

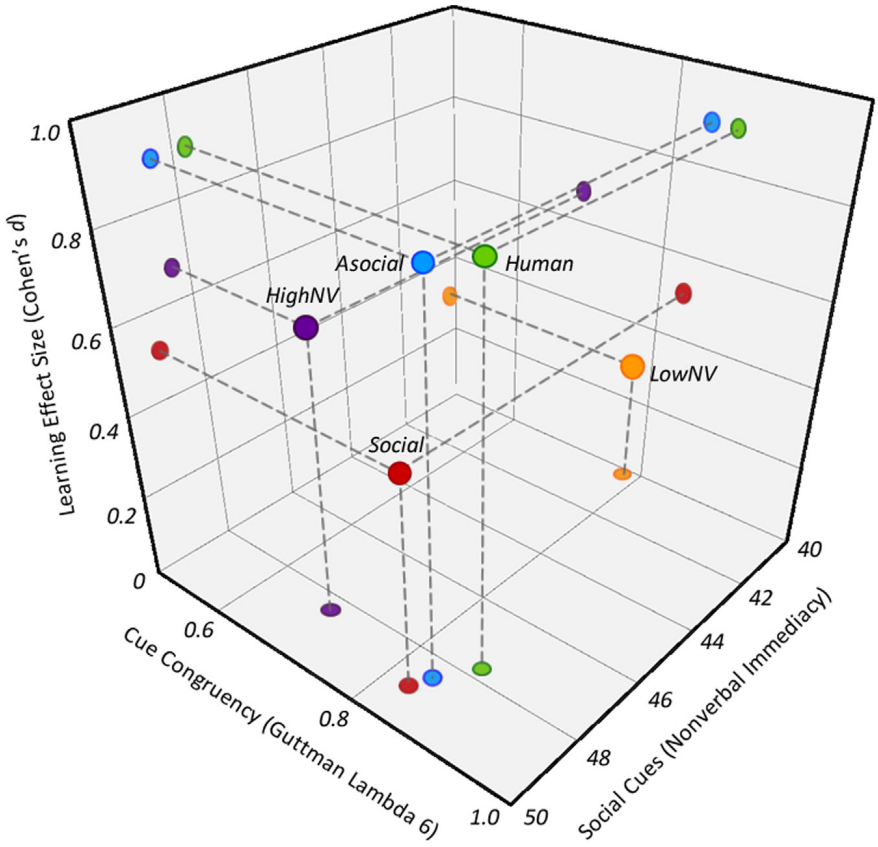


FIGURE 6 | Learning, congruency, and social behavior for each of the 5 conditions. Learning is measured in effect size between pretest and posttest for children. Congruency is indicated through Guttman's λ_6 of the adult nonverbal immediacy scores. Social behavior is characterized through nonverbal immediacy ratings from adults. An interactive version of this figure is available online to provide different perspectives of the space: https://goo.gl/ZNPxc8.

(indicated by Guttman's λ_6), learning also increases, and the same is true for social cues. The combination of congruency and social behavior as characterized by nonverbal immediacy provides a basis for learning predictions, where the combination of high social behavior and social cue congruency is necessary to maximize potential learning.

Such a hypothesis is supported by the view of social cues being perceived as a single percept, as suggested by Zaki (2013).

Experimental evidence with perception of emotions would seem to provide additional weight to such a perspective (Nook et al., 2015). This has clear implications for designers of social robot behavior when human perceptions or outcomes are of any degree of importance. The combination of all social cues in context must be considered alongside the expectations of the human to generate appropriate behavior. Not only does this give rise to a number of challenges, such as identifying combinatorial contextual expectations for social cues, but it could also have implications for how social cues should be examined experimentally. The isolation of specific social cues in experimental scenarios would not describe the role of that social cue, but the role of that social cue, *given the context of all other cues*. This is an important distinction that leads to a great deal more complexity in "solving" behavioral design for social robots, but that would also contribute to explanations of why a complex picture is emerging in terms of the effect of robot behavior on learning, as discussed in Section 2.1. The NVI metric and the predictions (that can be objectively examined) we put forward below provide a means through which robot behavior designers can iteratively implement and evaluate holistic social behaviors in an efficient manner, contributing to a more coherent framework in this regard. In particular, three predictions can be derived from the extremities of the space that is presented:

P1. Highly social behavior of a tutor robot (as characterized by nonverbal immediacy) with high congruency will lead to maximum potential learning.
P2. Low social behavior of a tutor robot with low congruency will lead to minimal potential learning.
P3. A mismatch in the social behavior of a tutor robot and the social cue congruency will lead to less than maximum potential learning.

Guttman's Lambda, as providing a measure of consistency, is used here as a proxy for the congruency of cues as observed by the study participants. We argue that this provides the necessary insight into cue congruency; however, the mapping between this metric and overtly judged congruency remains to be characterized. This would not necessarily be something that would be straightforward to achieve due to the potentially complex interactions between large numbers of social cues. For these predictions, use of the NVI metric as the characterization of social behavior would still suffer from some of the issues outlined earlier in this discussion: lack of timing information, relative cue importance, and novelty of behavior. The predictions are based on the general trends observed here, and it is noted that NVI is not a comprehensive measure of social behavior; indeed the SR condition in particular would not be fully explained using this means alone when compared to other results such as the AR condition. In addition, the data used for the learning axis were collected with relatively few samples (just over 10 per condition) in a specific experimental setup. Ideally, many further samples would be collected in both short and long term. The data collected here are over the short term and with children unfamiliar with robots. As longer term interactions take place, or as robots become more commonplace in society, expectations may change.

6. CONCLUSION

In this article, we have considered the use of nonverbal immediacy as a means of characterizing nonverbal social behavior in human–robot interactions. In a one-to-one maths tutoring task with humans and robots, it was shown that children and adults provide strong positively correlated ratings of tutor nonverbal immediacy. In addition, in agreement with the human–human literature, a positive correlation between tutor nonverbal immediacy and child learning was found. However, nonverbal immediacy alone could not account for all of the learning differences between tutoring conditions. This discrepancy led to the consideration of social cue congruency as an additional factor to social behavior in learning outcomes. Guttman's λ_6 was used to provide an indication of congruency between social cues. The combination of social behavior (as measured through nonverbal immediacy) and cue congruency (as indicated by Guttman's λ_6) provided an explanation of the learning data. It is suggested that if we are to achieve desirable outcomes with, and reactions to, social robots, greater consideration must be given to all cues in the context of multimodal social behavior and their possible perception as a unified construct. The hypotheses we have generated predict that the combination of high social behavior, and social cue congruency is necessary to maximize learning. The Robot Nonverbal Immediacy Questionnaire (RNIQ) developed for use here is offered as a means of gathering data for such characterizations.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Plymouth University ethics board with written informed consent from all adult subjects. Child subjects gave verbal informed consent themselves, and written informed consent was provided by a parent or guardian. All subjects gave informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Plymouth University ethics board.

AUTHOR CONTRIBUTIONS

Conception and design of the work, interpretation and analysis of the data, and draft and critical revisions of the work: JK, PB, and TB. Acquisition of the data: JK.

FUNDING

This work is partially funded by the EU H2020 L2TOR project (grant 688014), the EU FP7 DREAM project (grant 611391), and the School of Computing, Electronics and Maths, Plymouth University, UK.

REFERENCES

Alemi, M., Meghdari, A., and Ghazisaedy, M. (2014). Employing humanoid robots for teaching English language in Iranian junior high-schools. *Int. J. Hum. Robot.* 11, 1450022-1–1450022-25. doi:10.1142/S0219843614500224

Ambady, N., and Rosenthal, R. (1993). Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *J. Pers. Soc. Psychol.* 64, 431. doi:10.1037/0022-3514.64.3.431

Bartneck, C., Kanda, T., Mubin, O., and Al Mahmud, A. (2009a). Does the design of a robot influence its animacy and perceived intelligence? *Int. J. Soc. Robot.* 1, 195–204. doi:10.1007/s12369-009-0013-7

Bartneck, C., Kuli, D., Croft, E., and Zoghbi, S. (2009b). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.* 1, 71–81. doi:10.1007/s12369-008-0001-3

Baxter, P., Wood, R., and Belpaeme, T. (2012). "A touchscreen-based 'sandtray' to facilitate, mediate and contextualise human-robot social interaction," in *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: ACM), 105–106.

Belpaeme, T., Baxter, P., De Greeff, J., Kennedy, J., Read, R., Looije, R., et al. (2013). "Child-robot interaction: perspectives and challenges," in *Proceedings of the 5th International Conference on Social Robotics ICSR*', Vol. 13 (Cham, Switzerland: Springer), 452–459.

Borgers, N., Sikkel, D., and Hox, J. (2004). Response effects in surveys on children and adolescents: the effect of number of response options, negative wording, and neutral mid-point. *Qual. Quant.* 38, 17–33. doi:10.1023/B:QUQU.0000013236.29205.a6

Christensen, L. J., and Menzel, K. E. (1998). The linear relationship between student reports of teacher immediacy behaviors and perceptions of state motivation, and of cognitive, affective, and behavioral learning. *Commun. Educ.* 47, 82–90. doi:10.1080/03634529809379112

Comstock, J., Rowell, E., and Bowers, J. W. (1995). Food for thought: teacher nonverbal immediacy, student learning, and curvilinearity. *Commun. Educ.* 44, 251–266. doi:10.1080/03634529509379015

Cronbach, L. J., and Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educ. Psychol. Meas.* 64, 391–418. doi:10.1177/0013164404266386

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., and Wagner, S. (2001). Explaining math: gesturing lightens the load. *Psychol. Sci.* 12, 516–522. doi:10.1111/1467-9280.00395

Goldin-Meadow, S., Wein, D., and Chang, C. (1992). Assessing knowledge through gesture: using children's hands to read their minds. *Cogn. Instr.* 9, 201–219. doi:10.1207/s1532690xci0903_2

Gordon, G., Breazeal, C., and Engel, S. (2015). "Can children catch curiosity from a social robot?" in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: ACM), 91–98.

Gorham, J. (1988). The relationship between verbal teacher immediacy behaviors and student learning. *Commun. Educ.* 37, 40–53. doi:10.1080/03634528809378702

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika* 10, 255–282. doi:10.1007/BF02288892

Ham, J., Bokhorst, R., and Cabibihan, J. (2011). "The influence of gazing and gestures of a storytelling robot on its persuasive power," in *International Conference on Social Robotics* (Cham, Switzerland).

Han, J., Jo, M., Park, S., and Kim, S. (2005). "The educational use of home robots for children," in *Proceedings of the IEEE International Symposium on Robots and Human Interactive Communications RO-MAN*, Vol. 2005 (Piscataway, NJ: IEEE), 378–383.

Herberg, J., Feller, S., Yengin, I., and Saerbeck, M. (2015). "Robot watchfulness hinders learning performance," in *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, Vol. 2015 (Piscataway, NJ), 153–160.

Huang, C.-M., and Mutlu, B. (2013). "Modeling and evaluating narrative gestures for humanlike robots," in *Proceedings of the Robotics: Science and Systems Conference, RSS*' (Berlin), 13.

Jung, Y., and Lee, K. M. (2004). "Effects of physical embodiment on social presence of social robots," in *Proceedings of the 7th Annual International Workshop on Presence* (Valencia, Spain), 80–87.

Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: a field trial. *Hum. Comput. Interact.* 19, 61–84. doi:10.1207/s15327051hci1901&2_4

Kanda, T., Shimada, M., and Koizumi, S. (2012). "Children learning with a social robot," in *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: ACM), 351–358.

Kelley, D. H., and Gorham, J. (1988). Effects of immediacy on recall of information. *Commun. Educ.* 37, 198–207. doi:10.1080/03634528809378719

Kennedy, J., Baxter, P., and Belpaeme, T. (2015a). "Can less be more? The impact of robot social behaviour on human learning," in *Proceedings of the 4th International Symposium on New Frontiers in HRI at AISB 2015* (Canterbury).

Kennedy, J., Baxter, P., and Belpaeme, T. (2015b). 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

Kennedy, J., Baxter, P., and Belpaeme, T. (2015c). "The robot who tried too hard: social behaviour of a robot tutor can negatively affect child learning," in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: ACM), 67–74.

Kennedy, J., Baxter, P., Senft, E., and Belpaeme, T. (2015d). "Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions," in *Proceedings of the International Conference on Social Robotics* (Cham, Switzerland).

Kennedy, J., Baxter, P., and Belpaeme, T. (2017). Nonverbal immediacy as a characterisation of social behaviour for human-robot interaction. *Int. J. Soc. Robot.* 9, 109–128. doi:10.1007/s12369-016-0378-3

Kennedy, J., Baxter, P., Senft, E., and Belpaeme, T. (2016). "Heart vs hard drive: children learn more from a human tutor than a social robot," in *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction* (Piscataway, NJ: ACM), 451–452.

Leyzberg, D., Spaulding, S., Toneva, M., and Scassellati, B. (2012). "The physical presence of a robot tutor increases cognitive learning gains," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society, CogSci*, Vol. 2012 (Austin, TX), 1882–1887.

McCroskey, J. C., Sallinen, A., Fayer, J. M., Richmond, V. P., and Barraclough, R. A. (1996). Nonverbal immediacy and cognitive learning: a cross-cultural investigation. *Commun. Educ.* 45, 200–211. doi:10.1080/03634529609379049

Mehrabian, A. (1968). Some referents and measures of nonverbal behavior. *Behav. Res. Methods Instrum.* 1, 203–207. doi:10.3758/BF03208096

Moore, R. K. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Nat. Sci. Rep.* 2:864. doi:10.1038/srep00864

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* 19, 98–100. doi:10.1109/MRA.2012.2192811

Moshkina, L., Trickett, S., and Trafton, J. G. (2014). "Social engagement in public places: a tale of one robot," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: ACM), 382–389.

Nook, E. C., Lindquist, K. A., and Zaki, J. (2015). A new look at emotion perception: concepts speed and shape facial emotion recognition. *Emotion* 15, 569–578. doi:10.1037/a0039166

Nussbaum, J. F. (1992). Effective teacher behaviors. *Commun. Educ.* 41, 167–180. doi:10.1080/03634529209378878

Reeves, B., and Nass, C. (1996). *How People Treat Computers, Television, and New Media like Real People and Places*. New York, NY: CSLI Publications, Cambridge University press.

Revelle, W., and Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: comments on Sijtsma. *Psychometrika* 74, 145–154. doi:10.1007/s11336-008-9102-z

Richmond, V., McCroskey, J., and Payne, S. (1987). *Nonverbal Behavior in Interpersonal Relations*. Englewood Cliffs, NJ: Prentice-Hall.

Richmond, V. P., McCroskey, J. C., and Johnson, A. D. (2003). Development of the Nonverbal Immediacy Scale (NIS): measures of self- and other-perceived nonverbal immediacy. *Commun. Q.* 51, 504–517. doi:10.1080/01463370309370170

Robinson, R. Y., and Richmond, V. P. (1995). Validity of the verbal immediacy scale. *Commun. Res. Rep.* 12, 80–84. doi:10.1080/08824099509362042

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, CHI'10 (New York, NY: ACM), 1613–1622.

Sung, J., Christensen, H. I., and Grinter, R. E. (2009). "Robots in the wild: understanding long-term use," in *Prcoeedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY: IEEE), 45–52.

Szafir, D., and Mutlu, B. (2012). "Pay attention! Designing adaptive agents that monitor and improve user engagement," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'12* (New York, NY: ACM), 11–20.

Tanaka, F., and Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: field experiments in a classroom for vocabulary learning. *J. Hum. Robot Interact.* 1, 78–95. doi:10.5898/JHRI.1.1.Tanaka

Wainer, J., Feil-Seifer, D. J., Shell, D. A., and Mataric, M. J. (2007). "Embodiment and human-robot interaction: a task-based perspective," in *Proceedings of the 16th IEEE International Symposium on Robot and Human interactive Communication (IEEE), RO-MAN*, Vol. 2007 (Piscataway, NJ), 872–877.

Wilson, J. H., and Locker, L. Jr. (2007). Immediacy scale represents four factors: nonverbal and verbal components predict student outcomes. *J. Classroom Interact.* 42, 4–10.

Witt, P. L., Wheeless, L. R., and Allen, M. (2004). A meta-analytical review of the relationship between teacher immediacy and student learning. *Commun. Monogr.* 71, 184–207. doi:10.1080/036452042000228054

Zajonc, R. B. (1965). Social facilitation. *Science* 149, 269–274. doi:10.1126/science.149.3681.269

Zaki, J. (2013). Cue integration a common framework for social cognition and physical perception. *Perspect. Psychol. Sci.* 8, 296–312. doi:10.1177/1745691613475454

Conflict of Interest Statement: 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.

APPENDIX

A. Robot Nonverbal Immediacy Questionnaire (RNIQ)

The following is the questionnaire used by participants in the evaluation to rate the nonverbal immediacy of the robot, based on the short-form nonverbal immediacy scale-observer report. Options are provided in equally sized boxes below each question (or equally spaced radio buttons in the online version). The options are: **1 = Never; 2 = Rarely; 3 = Sometimes; 4 = Often; 5 = Very Often**. The questions are as follows:

1. The robot uses its hands and arms to gesture while talking to you
2. The robot uses a dull voice while talking to you
3. The robot looks at you while talking to you
4. The robot frowns while talking to you
5. The robot has a very tense body position while talking to you
6. The robot moves away from you while talking to you
7. The robot changes how it speaks while talking to you
8. The robot touches you on the shoulder or arm while talking to you
9. The robot smiles while talking to you
10. The robot looks away from you while talking to you
11. The robot has a relaxed body position while talking to you
12. The robot stays still while talking to you
13. The robot avoids touching you while talking to you
14. The robot moves closer to you while talking to you
15. The robot looks keen while talking to you
16. The robot is bored while talking to you

Scoring:

Step 1. Add the scores from the following items: 1, 3, 7, 8, 9, 11, 14, and 15.

Step 2. Add the scores from the following items: 2, 4, 5, 6, 10, 12, 13, and 16.

Total Score = 48 plus Step 1 minus Step 2.

Comparing L2 Word Learning through a Tablet or Real Objects: What Benefits Learning Most?

Rianne Vlaar *
Utrecht University, Netherlands

Josje Verhagen
Utrecht University, Netherlands

Ora Oudgenoeg-Paz
Utrecht University, Netherlands

Paul Leseman
Utrecht University, Netherlands

ABSTRACT

In child-robot interactions focused on language learning, tablets are often used to structure the interaction between the robot and the child. However, it is not clear how tablets affect children's learning gains. Real-life objects are thought to benefit children's word learning, but it is not clear whether virtual objects provide the same learning experiences. The present study aims to find out whether there is a difference in L2 vocabulary learning gains between children who manipulate physical objects and children who manipulate 3D models of the same objects on a tablet screen during a word-learning task. Data indicate no clear benefit of real-life objects over virtual objects.

Keywords

L2 word learning; Child-robot interaction; Embodiment; Tablets; Real-life objects

ACM Reference format:
R. Vlaar, J. Verhagen, O. Oudgenoeg-Paz, and P.P.M. Leseman. 2017. Comparing L2 Word Learning through a Tablet or Real Objects: What Benefits Learning Most?. In *Proceedings of ACM HRI conference, Vienna, Austria, March 2017 (R4L workshop at HRI 2017),* 2 pages.
DOI: 10.475/123 4

1. INTRODUCTION

In recent years, robots have been employed more and more for language tutoring purposes. In many of these child-robot interactions, a tablet is used to establish common ground and to ensure a successful interaction between the robot and the child [4,6]. However, it is not clear how the use of tablets in these interactions affects learning gains.

* Corresponding author: m.a.j.vlaar@uu.nl

R4L workshop at HRI 2017, Vienna, Austria

DOI: 10.475/123 4

The embodied-cognition approach, which states that language is grounded in real-life sensorimotor interactions [3], predicts that children's interactions with real-life objects benefit vocabulary learning [2,5]. From this approach, one would expect children to learn new words better if they manipulate physical objects rather than virtual objects on a tablet, as the former allow children to experience sensorimotor interactions with the objects. It is not yet clear, however, whether this actually is the case. Here, we report data from an experiment comparing the effect of real objects versus virtual objects on a tablet screen on L2 word learning. The main research question is whether there is a difference in L2 vocabulary learning gains between children who manipulate physical objects and children who manipulate 3D models of the same objects on a tablet screen. This question is not only relevant for language-learning theories, but to the field of robotics as well, for its implications on the design of robot-assisted language learning tasks.

2. PRESENT EXPERIMENT

Participants: Forty-six Dutch kindergartners (M = 60.6 months, age range = 50-73 months, SD = 6.77; 26 girls) with no knowledge of English participated in the experiment. Most children had experience working with touch screens, and all practiced with the tablet prior to the training.

Procedure: A pre-test was used to make sure the children did not know the target words. The training immediately followed the pre-test, using a between-subjects design such that children were randomly assigned to either the tablet or objects condition (n = 25 in the tablet condition; n = 21 in the object condition). Various tests were administered to measure the children's knowledge of the target words. One week later, the same tests were re-administered to measure children's retention of the target words.

Materials: In the training, children were presented with a story in Dutch containing six L2 (English) target words (i.e., 'heavy', 'light', 'full', 'empty', 'in front of,' and 'behind'). These targets were chosen as children should benefit from sensorimotor interactions with objects in learning them. For example, learning the word "heavy" could be easier when actually holding a heavy object than seeing a 3D model of this object on a tablet screen. The target

words were each presented ten times. During the training, children were asked to repeat each target word once, translate the Dutch word to its English equivalent, and perform simple actions in relation to these words on either the tablet or with the real objects (e.g., put a "heavy" elephant in its cage).

The immediate and delayed post-tests included several tasks to assess children's learning of the L2 words. Two translation tasks (English to Dutch and Dutch to English; maximum score six) were used to measure productive vocabulary. To measure receptive vocabulary, a comprehension task in which children were asked to select the picture (out of four options) which best matched the target words (maximum score twenty-four), and a sorting task was used in which children had to sort pictures in trays according to their meaning, per word pair of antonyms (i.e., all the "heavy" pictures in one tray; all the "light" pictures in the other tray; maximum score thirty). Last, a story comprehension task was used to measure the child's recall of the narrative (maximum score six).

3. RESULTS

Independent-samples *t*-tests revealed no significant differences between using a tablet or physical objects on any of the tasks, as indicated by children's mean accuracy scores on the direct and delayed post-tests (see Figure 1 and 2; all $ps > .243$). In the receptive tests (the comprehension task and sorting task), children scored significantly above chance level (indicated by the black line), irrespective of condition (all $ps < .001$). In the production test (the translation tasks), children accurately produced one or two translations. Children also showed proper recall of the narrative, as indicated by the data of the story task in both conditions. Interestingly, in both conditions, the mean scores on the Dutch-to-English translation task were higher for the delayed post-test than for the immediate post-test (both $ps < .001$), possibly indicating some sort of sleep effect (see [1] for an overview).

4. DISCUSSION

The data show that children's manipulations of physical objects or virtual objects on a tablet screen do not affect L2 vocabulary learning gains differently. These results may be due to the fact that we studied L2 word learning as opposed to L1 learning. In L1 word learning, one has to learn both the word form and the concept, while in L2 learning, one can often make use of the L1 knowledge and connect it to the L2 word form. It is possible sensorimotor interactions with objects do not affect learning gains as much when one has already acquired a concept in

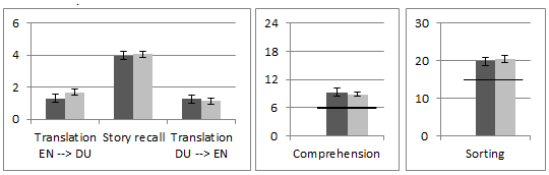


Figure 1. Mean accuracy scores on the direct post-test (dark grey = object condition; light grey = tablet condition)

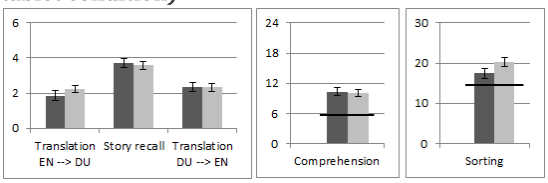


Figure 2. Mean accuracy scores on the delayed post-test (dark grey = object condition; light grey = tablet condition)

their L1, and can subsequently use this knowledge in learning the L2 word.

Future research should therefore look into L1 word learning with objects or tablets, or L2 words of which the concepts do not match the L1 concept the child has acquired. However, present data indicate virtual objects on a tablet screen can be incorporated in child-robot interaction studies on L2 vocabulary learning.

5. ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 688014.

REFERENCES

[1] E. L. Axelsson, S.E. Williams, S. E., and J.S. Horst. The Effect of Sleep on Children's Word Retention and Generalization. *Frontiers in psychology*, 7, 2016.

[2] A.M. Glenberg. Embodiment for education. In: P. Calvo, & T. Gomila (Eds.). *Handbook of Cognitive science: An embodied approach*. Amsterdam, The Netherlands: Elsevier. 2008.

[3] S.A. Hockema and L.B. Smith. Learning your language, outside-in and inside-out. *Linguistics,* 47: 453-479, 2009.

[4] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme. Social robot tutoring for child second language learning. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, pages 67-74. ACM, 2016.

[5] A.W. Kersten and L.B. Smith. Attention to novel objects during verb learning. *Child Development*, 73: 93-109. 2002

[6] J. Kory Westlund and C. Breazeal. The interplay of robot language level with children's language learning during storytelling. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction,* pages 65-66. ACM, 2015.

Exploring the Effect of Gestures and Adaptive Tutoring on Children's Comprehension of L2 Vocabularies

Jan de Wit
TiCC*
Tilburg University
j.m.s.dewit@uvt.nl

Thorsten Schodde
Faculty of Technology, CITEC‖
Bielefeld University
tschodde@techfak.uni-bielefeld.de

Bram Willemsen
TiCC*
Tilburg University
b.willemsen@uvt.nl

Kirsten Bergmann
Faculty of Technology, CITEC‖
Bielefeld University
kirsten.bergmann@uni-bielefeld.de

Mirjam de Haas
TiCC*
Tilburg University
mirjam.dehaas@uvt.nl

Stefan Kopp
Faculty of Technology, CITEC‖
Bielefeld University
skopp@techfak.uni-bielefeld.de

Emiel Krahmer
TiCC*
Tilburg University
e.j.krahmer@uvt.nl

Paul Vogt
TiCC*
Tilburg University
p.a.vogt@uvt.nl

ABSTRACT

The L2TOR project explores the use of social robots for second language tutoring. This paper presents an experiment in preparation to investigate the effects of two educational scaffolding features (adaptation/personalization and iconic gestures), when used by a robot tutor, on children's comprehension of animal names in a foreign language. Participants will be children between the ages of four and five. The study is scheduled to take place in March 2017.

CCS CONCEPTS

•Computing methodologies → Cognitive robotics; Probabilistic reasoning; •Applied computing → Interactive learning environments; •Human-centered computing → *Empirical studies in HCI;*

KEYWORDS

Language tutoring; Assistive robotics; Education; Bayesian knowledge tracing; Human-robot interaction

ACM Reference format:
Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2017. Exploring the Effect of Gestures and Adaptive Tutoring on Children's Comprehension of L2 Vocabularies. In *Proceedings of ACM HRI conference, Vienna, Austria, March 2017 (R4L workshop at HRI 2017),* 6 pages.
DOI: 10.475/123_4

1 INTRODUCTION

The L2TOR project aims to design and develop a robot tutor capable of supporting children of four to five years old in the acquisition of a second language by interacting naturally with them in their social and referential environment through one-to-one tutoring interactions [1]. The robot used for the L2TOR project is the SoftBank Robotics NAO humanoid robot. The NAO robot is capable of speaking multiple languages, readily able to switch between them, which provides the possibility to vary the amount of the child's native language (L1) and the second language (L2) to be taught. Furthermore, the physical presence of a robot is shown to improve learning gains compared to its two-dimensional counterparts (e.g. Leyzberg et al. [12]).

This three-year project will result in an integrated lesson plan, which is expected to contain 24 lessons spanning three different domains (math, space, and mental state). To design these lessons, we analyze the way human tutors interact with children and investigate how different functionalities of the robot can be used to ensure a natural and productive interaction. In this paper, we propose an experiment to evaluate two such functionalities: personalized lessons by adjustment of the level of difficulty of the subject matter to the level of proficiency of the learner and the use of gestures when introducing the L2 words. We expect that both concepts will help to create and maintain common ground with the child, while also increasing comprehension and memorization potential of new words in the L2.

The importance of personalized adjustments in the robot's behavior has been substantiated in recent research showing that participants who received personalized lessons from a robot (based on heuristic skill assessment) outperformed others who received a non-personalized training [12]. Suboptimal robot behavior (e.g. distracting, incongruent or in other ways inappropriate social behavior) can even hamper learning [10].

One of the main advantages of choosing a humanoid robot as a tutor is its physical presence in the world, allowing for interactions similar to those between humans. Because of its anthropomorphic appearance, we tend to expect human-like communicative behavior

R4L workshop at HRI 2017, Vienna, Austria

DOI: 10.475/123_4

*Tilburg center for Cognition and Communication
‖Cluster of Excellence Cognitive Interaction Technology

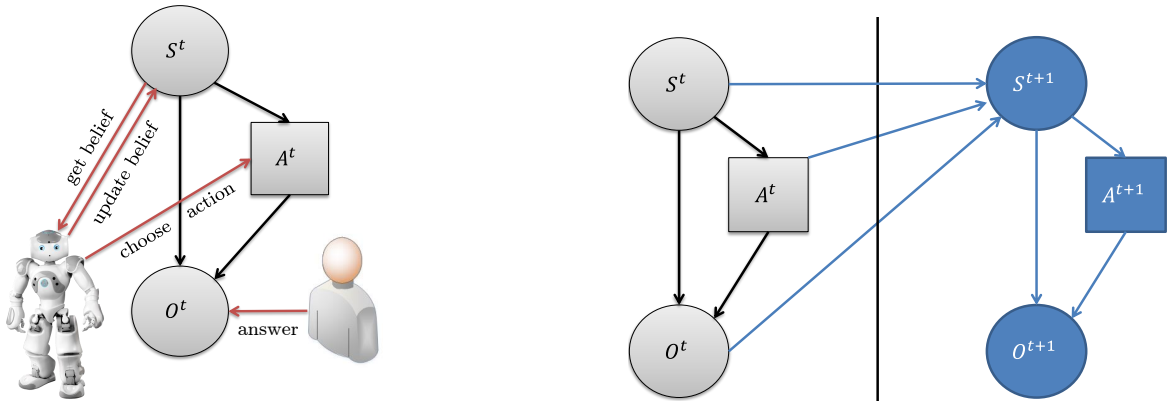


Figure 1: Dynamic Bayesian Network for BKT: With the current skill-belief the robot chooses the next skill S^t and action A^t for time step t (left). After observing an answer O^t from the learner, this observation together with action A^t and the previous skill-belief S^t are used to update the skill-belief S^{t+1} at time $t+1$ (right) [18].

from the robot, including proper use of non-verbal communication. Robots that perform gestures are perceived in a more positive way than those that use only speech [16].

In Section 2 we explain our previous work to evaluate adaptive learning, which is used as a starting point for the experiment described in this paper. We then introduce iconic gestures and describe how they could be used to increase learning gain in a human-robot tutoring context in Section 3, followed by our main research questions in Section 4. Section 5 outlines the design of the proposed experiment. We intend to start data collection in March 2017.

2 PREVIOUS WORK

2.1 Adaptive language tutoring with a robot

In previous work we developed a novel approach to personalize language tutoring in human-robot interaction [18]. This adaptive tutoring is enabled through a model of how tutors mentalize about learners – by keeping track of their knowledge state and by selecting the next tutoring actions based on their likely effects on the learner. This is realized via an extended model that combines knowledge tracing (of what the learner learned) with tutoring actions (of the tutor) in one causal probabilistic model. This allows for selecting skills and actions based on notions of optimality – here the desired learner's knowledge state as well as optimal task difficulty – to achieve this for a given skill.

The approach is based on Bayesian Knowledge Tracing (BKT) [4], a specific type of Dynamic Bayesian Networks (DBNs). The model consists of two types of variables, namely the *latent variables* representing the belief state of 'skills' to be acquired (e.g. whether a word has been learned or not) and the *observed variables* representing the observable information of the learning interaction (e.g. whether an answer was correct or not). In our proposed model, each latent variable can attain six discrete values, corresponding to six bins for the belief state (0%, 20%, 40%, 60%, 80%, 100%) representing whether a skill is mastered or not as a discretized probability distribution. That is, we reduce the complexity we would get through continuous latent variables but also attain more flexibility. The observed variables remain binary and still represent whether a learner's response is correct or not (see Figure 1). Moreover, the following update of the belief state of the skill, i.e. the skill-belief, at time $t+1$ is not only based on the previous skill-belief, but also on the chosen action and the previous observation at time t.

Based on this model, two types of decisions are made, (1) which skill would be the best to address next, and (2) the choice of action to address that skill. Regarding the former, we employ a heuristic maximizing the beliefs of all skills while balancing the single skill-beliefs among each other. This strategy is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system [11]. Regarding the choice of action, the model enables the simulation of the impact each action has on a particular skill. To keep the model simple, the action space of the model consists of three different task difficulties (easy, medium, hard). Consider an example where the skill-belief appears relatively high, such that the skill is nearly mastered by the learner. In this case, a less challenging task would only result in a relatively minor benefit for the training of that skill. In contrast, if we assume the skill-belief to be rather low and a very difficult task is given, the student would barely be able to solve the task, likewise resulting in a smaller (or non-existent) learning gain. Instead, a task of adequate difficulty, not needlessly simple nor too complicated for the student to solve, will result in a higher learning gain [5]. This helps to position the robot as a capable instructor that uses these scaffolding techniques to help children acquire new skills beyond what they could have learned without help, by bringing them into the zone of proximal development (ZPD) [22].

2.2 Evaluation

When implemented in the robot language tutor, the model will enable the robot tutor to trace the learner's knowledge with respect to the words to be learned, to decide which skill (word) to teach next, and how to address the learning of this skill in a game-like tutoring interaction. For the experiment as described in [18], participants were asked to learn ten vocabulary items German – 'Vimmi'

(Vimmi is an artificial language that was developed to avoid associations with other known words or languages for language-related experiments [13]). The items included colors, shapes and the words 'big' and 'small'. During the game, the robot would introduce one of the Vimmi words. A tablet then displayed several images, one of which satisfied the Vimmi description (e.g. one object that is blue) and a number of distractors. The participant was then asked to select the image corresponding to the described item. Participants learned vocabulary items in one of two conditions, either in the condition with the adaptive model or in a non-adaptive (random) control condition. In the adaptive condition, the skill to be taught and the action to address the skill were chosen by the model as described above. Participants' performance was assessed with two measures: (1) learners' response behavior was tracked over the course of the training to investigate the progress of learning, and (2) a post-test was conducted on the taught vocabulary in the form of both L1-to-L2 translations and L2-to-L1 translations to assess participants' state of knowledge following the intervention.

Analysis of participants' response behavior over the course of training indicated that the participants learned the L2 words during the human-robot interaction (see [18] for more detailed results). Importantly, they learned more successfully with our adaptive model as compared to a randomized training. That is, the repeated trials addressing still unknown items as chosen by the adaptive model (until the belief state about these words equaled that of known items) outperformed the tutoring of the same material (same number of trials and items) but in randomized order. In the post-test, however, there was no significant difference across experimental conditions, despite a trend towards increased performance in the adaptive model conditions as compared to the controls.

3 ICONIC GESTURES

A growing body of evidence suggests that iconic gestures bear a great potential to enhance learners' memory performance for novel L2 words. Iconic gestures are movements that have a formal relation (in form or manner of execution) to the semantic content of the linguistic unit they describe [14]. In other words, the gesture elicits a mental image that relates strongly to the word or words that it links to. As an example, the word *bird* could be described by an iconic movement of stretching both arms sideways and moving them up and down, symbolizing the flapping of wings. The supporting effect of iconic gestures on L2 vocabulary learning by providing a congruent link between the words to be learned and gesture being observed or imitated has been shown in various studies (e.g. [6, 9, 13, 15, 19]). A recent overview of how gestures contribute to foreign language learning and possible explanations for this effect is given by Hald et al. [8]. Although they focus mainly on students *performing* or re-enacting the gestures, merely observing a gesture is shown to aid learning as well. Research conducted by Tellier [19] and De Nooijer et al. [6] investigated the role of gestures with respect to children and word learning. The effect of gestures is shown to depend on the students' gender, language background and existing experience with the L1 [15].

When considering the use of an artificial embodied agent as a tutor, the positive effects of gesturing seem to apply as well, as shown by Bergmann and Macedonia for a virtual tutor [2], and by Van Dijk et al. for a robotic tutor [20]. An additional benefit of implementing non-verbal behavior is to improve the way the robot is perceived, making it seem more human-like[17]. The challenge of mapping non-verbal behavior to the robot lies in the fact that each act needs to be carefully designed and choreographed to coincide with the corresponding word or sentence. There are limits to the degrees of freedom, the working space (i.e. the physical reach) and smoothness of motion that the robot has to offer. As an example, Figure 2 shows an attempt at making an iconic gesture for *rabbit*. The right arm has to take an unnatural position, which may result in an uncanny feeling for the observer. The NAO robot also has only three fingers and they cannot move independently, therefore finger-counting and similar subtle motions do not transfer to the robot without modification. The challenge lies in finding ways to work around these limitations, while still taking advantage of the added value of non-verbal communication. The gestures that were designed for this experiment have been exaggerated beyond what the human alternatives would look like. For example, when imitating a monkey the robot will bend its knees and shift its weight from side to side (see Figure 3).

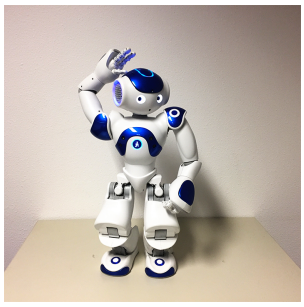


Figure 2: Attempt at showing an iconic gesture for a *rabbit*. The unnatural angle of the arm, positioning of the hand, and movement of the fingers, may lead to confusion and, consequently, adverse effects with respect to learning.

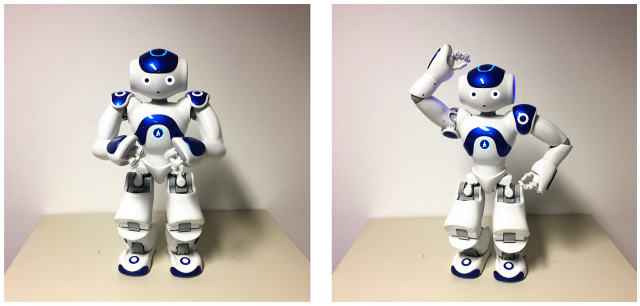


Figure 3: Stills of iconic gestures as depicted by the robot. Left: imitating a *chicken* by simulating the flapping of its wings; right: imitating a *monkey* by simulating the scratching of the head and armpit with the right and left extremities, respectively.

4 RESEARCH QUESTIONS

With the upcoming experiment we intend to answer two research questions. The first question relates to the previous work described in Section 2. We aim to investigate to what extent children will benefit from adaptive language tutoring. We hypothesize an increase in learning gain when children are taught words through an adaptive language tutoring system as compared to a non-adaptive (random) language tutoring system. We anticipate a difference in the exact words that are learned: in the adaptive condition, we expect children to learn those words that were the most challenging during training (having the most incorrect answers) because of the higher repetition rate of these words. In the random condition, the words learned might depend on other factors such as word complexity or attitude towards the animal described by the word.

Our second research question focuses on the effect of gestures on L2 comprehension for children. We hypothesize an increase in learning gain when target words are accompanied by (iconic) gestures during learning, as compared to the absence of gestures. Furthermore, we expect a reduced knowledge decay over time of the words in the gesture condition, similar to the discoveries by Cook et al. in the math problem solving domain with a human tutor [3]. We intend to investigate, using the retention test one week after the experiment, whether these findings carry over to the language learning domain with gestures performed by the robot. It should be noted that participants are not required but also not prohibited from using gestures during the experiment and pre- and post-tests. We are interested in seeing whether children will produce gestures spontaneously following training and, if so, to what extent these gestures will prove to be similar to the ones depicted by the robot.

5 PROPOSED EXPERIMENT

Following the two research questions, our experiment has a 2 (adaptive versus non-adaptive) x 2 (gestures versus no gestures) between-subjects design. We aim to recruit 80 participants, all native Dutch speaking children between the ages of four and five.

Although the proposed experiment is largely a replication of the experiment described in Section 2 and presented in [18], changes to the design had to be made to accommodate the younger participants, as the previous experiment was tailored to adults. Instead of the first interaction between the children and the robot taking place as part of the experiment, the robot will be introduced to the children in a group session the week prior to the experiment to build trust and rapport. We will refer to the robot by a proper name (Robin) and present a background story to stimulate a friendly and open attitude towards the robot [21].

Rather than teaching children the fictional Vimmi words, the target words are the English names of six animals: chicken, monkey, horse, spider, bird, and hippo (used instead of the more difficult hippopotamus). The number of words was reduced to six (from ten in the original experiment, see Schodde et al. [18]) to account for the lower word memory span of children [7], which should be around four words for children of age five. All target words have been selected based on the (varying degrees of) dissimilarity between the words in the L1 (Dutch) and the L2 (English) as well as the feasibility of designing suitable iconic gestures to be performed by the robot to depict the animals. We will conduct a pre-test

Figure 4: Mock-up of the training phase of the proposed experiment. Three animals appear on the tablet screen, one of which matches the animal picked by the robot. The robot asks the child in their L1 to point out the correct animal based on its name in the L2. In the gesture condition, as shown in this image, the robot performs the associated iconic gesture when mentioning the animal.

to verify that participants are familiar with all six target words in their L1 (Dutch) and to test the underlying assumption that participants have no prior knowledge of the target words in the L2 (English). This pre-test will be presented on a different computer screen than the one on which the game is played and without the robot being present, so that there is a clear distinction between this testing environment and the training (game) stage. On the computer screen, the participant will be presented with the pictures of all six animals, one by one. For each picture, the experimenter will ask the participant for the name of the animal in the L1. The computer will then show the pictures of all animals on the screen and name the animals, one after another, in the L2 in random order. Each time the child is prompted with a name in the L2, they are asked to pick the correct image for this animal from the six animals displayed.

The experimental setup uses a Microsoft Surface Pro 4 tablet and the SoftBank Robotics NAO robot. The robot plays a game of "I spy with my little eye", where it picks a certain animal displayed on the tablet screen and names it in the L2, after which the child is expected to tap the corresponding animal picture (see Figure 4). The experimenter inputs the name of the child, so that the robot can personally address the participant, and starts the game. After a brief explanation, the tablet will ask participants to indicate whether they understand the concept of the game. If they indicate that they do not, the experimenter will intervene to provide further explanations. The experiment can be stopped at any time via an experimenter-controlled control panel. Once the actual game commences, the experimenter pretends to be preoccupied so as to avoid participants actively looking for feedback.

In the adaptive learning condition the next target word to train is selected based on the knowledge model (i.e. skill-beliefs) of the participant. After each trial in which the robot exposes the child to one animal, this knowledge model is updated based on the responses of the child. The updated model is then used to select the next target word to be presented. In the random condition, target

words are instead randomly presented. In total, there are thirty of these tasks, which means that in the random condition each target word is presented five times throughout the game. In the adaptive condition, the number of times each word occurs depends on how well the participant performs on that specific word, but all words are guaranteed to occur at least once. The previous experiment also consisted of a total of thirty tasks, but as there were ten target words there was less repetition. Reducing the number of words should avoid cognitive overload for the young participants while simultaneously offering more room for the adaptive system to learn the knowledge model of the child and repeat the words that require more attention.

A new addition to the experiment is a condition in which the robot will perform iconic gestures whenever one of the animal names is mentioned in English. These gestures were specifically designed for this experiment, where the robot tries to mimic the appearance or behavior of the animal. The timing of L2 word pronunciation is designed to occur close to the stroke of the gesture. This means that there is a pause in mid-sentence leading up to and after the L2 word, creating additional emphasis on the target. In the condition without gestures, a similar pause is introduced. The robot is set to "breathing mode" in all conditions, which means that it slowly moves its weight from one leg to the other while slightly moving its arms. This prevents the robot from being completely static while, in the gesture condition, reducing the surprise effect of an iconic gesture being triggered.

After thirty prompts to recognize the English animal names, the game finishes. The child is then presented with the post-test, again at the computer screen without the robot. The post-test is identical to the pre-test, except that we no longer test the animal names in the L1. The post-test is also identical across all conditions, so there are no gestures when the L2 words are presented. There are two different images for each animal, one of which will be used for the pre-test and post-test and the other for the game. The images of animals used in the pre-test and post-test feature the same character as those that appear during the game, but in a different pose. The pose in the set of images used during the game is designed to match the gesture that is shown by the robot, to avoid having a mismatch between both sources of visual information for some animal names, and a match for others [23]. For instance, for the word 'bird' the robot will display the act of flying by moving its arms up and down, therefore the bird in the image is also flying. The second set of images could feature the bird facing a different direction, sitting still. By using these two sets of images, we aim to test if children manage to map the English words not only to the specific corresponding image or mental representation of shape, but to the general concept of the animal. One week after the experiment we perform the post-test once again to measure the retention of the six target words.

To assess the iconicity of the gestures, we conducted a perception test with adult participants through an online survey. Participants ($N = 14$) were shown video recordings, one after another, of the six gestures performed by the robot. For each video, participants were asked to answer the question which animal the robot depicted by selecting the corresponding name of the animal in English from a list containing all six target words. The order in which the videos were shown, as well as the order of the items on the list containing the six animal names, was randomized for each participant. Results from the perception test are presented in Table 1. As can be seen from this confusion matrix, with an average accuracy of over 89 percent, participants were, on average, very accurate with respect to their predictions of the depicted gestures. In fact, for three of the six animals (monkey, horse, and bird), not a single mistake was made. With an average accuracy of just over 71 percent, the most ambiguous gestures were those representing the chicken and the hippo. However, it should be noted that participants typically came to realize they had made a mistake, after which they acted accordingly: for example, if a participant was shown the video recording of the chicken prior to that of the monkey and they had incorrectly selected *monkey* as their answer for the recording of the chicken, they would (now correctly) select *monkey* again as their answer when shown the recording of the monkey (we did not allow them to directly correct their past mistake). This implied correction, as well as the high accuracy on average, suggests that we may assume the gestures to be sufficiently iconic, especially as they will ultimately be presented in combination with the verbalization of the name of the associated animal.

Table 1: Confusion Matrix Perception Test

		Perceived					
		Chicken	Monkey	Horse	Spider	Bird	Hippo
Actual	Chicken	10	2	1	0	0	0
	Monkey	0	14	0	0	0	0
	Horse	0	0	14	0	0	0
	Spider	0	0	1	13	0	0
	Bird	0	0	0	0	14	0
	Hippo	1	1	0	2	0	10

Note. Shaded cells indicate true positives.

In our analysis of the experimental results, we intend to measure performance (correct and incorrect answers) during the word training to monitor participants' progress over time in the different conditions. Time on task is measured both in the training "game" and in the post-test. In addition, we will make video recordings of the interaction with the robot for additional analyses (for instance to see if and at what rate children will mimic the robot's gestures). During the post-test we will record how many animals the children managed to correctly identify immediately after training. The retention test will measure decay of the newly attained words after one week.

6 CONCLUSION

The experiment proposed in this paper outlines two valuable topics of discussion for improving the interactions between children and robot, specifically in a tutoring setting. We aim to investigate how the order and frequency of presenting new words in the L2 for the purpose of second language learning can be personalized for each child to optimize learning gain, based on a probabilistic model that traces their knowledge of each word. Second, the experiment

evaluates if the positive effect of performing iconic gestures for second language learning by human tutors carries over to the robot.

After running the experiment, future work includes incorporating our findings into the L2TOR project[1]. Adaptive learning will be integrated with the existing lesson plans, improving not only the way the content of each individual lesson is structured but also informing the choice of which words from previous lessons to repeat for better retention. If iconic gestures indeed prove to play a big part in learning and remembering new words, more of these non-verbal behaviors will be developed to accompany a greater number of (target) words and concepts. Furthermore, we will investigate the use of different types of gestures and explore ways of reducing the effort required to implement and orchestrate these gestures for robots. Our progress can be tracked via the project website[1].

7 ACKNOWLEDGMENTS

This work is partially funded by the H2020 L2TOR project (grant 688014), the Tilburg center for Cognition and Communication 'TiCC' at Tilburg University (Netherlands) and the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), funded by the German Research Foundation (DFG), at Bielefeld University (Germany). We would like to thank all members of the L2TOR project for their valuable comments and suggestions that have contributed towards the design of the experiment.

REFERENCES

[1] Tony Belpaeme, James Kennedy, Paul Baxter, Paul Vogt, Emiel E.J. Krahmer, Stefan Kopp, Kirsten Bergmann, Paul Leseman, Aylin C. Küntay, Tilbe Göksun, Amit K. Pandey, Rodolphe Gelin, Petra Koudelkova, and Tommy Deblieck. 2015. L2TOR - Second Language Tutoring using Social Robots. In *Proceedings of the International Conference on Social Robotics (ICSR) 2015 WONDER Workshop.*

[2] Kirsten Bergmann and Manuela Macedonia. 2013. *A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance.* Springer Berlin Heidelberg, Berlin, Heidelberg, 139–148. DOI: http://dx.doi.org/10.1007/978-3-642-40415-3_12

[3] Susan Wagner Cook, Zachary Mitchell, and Susan Goldin-Meadow. 2008. Gesturing makes learning last. *Cognition* 106, 2 (2008), 1047–1058.

[4] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* 4, 4 (1994), 253–278. DOI: http://dx.doi.org/10.1007/BF01099821

[5] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media* 29, 3 (2004), 241–250. DOI: http://dx.doi.org/10.1080/1358165042000283101

[6] Jacqueline A. de Nooijer, Tamara van Gog, Fred Paas, and Rolf A. Zwaan. 2013. Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. *Acta Psychologica* 144, 1 (2013), 173 – 179. DOI: http://dx.doi.org/10.1016/j.actpsy.2013.05.013

[7] Frank N. Dempster. 1981. Memory span: Sources of individual and developmental differences. *Psychological Bulletin* 89, 1 (1981), 63–100. DOI: http://dx.doi.org/10.1037/0033-2909.89.1.63

[8] Lea A. Hald, Jacqueline de Nooijer, Tamara van Gog, and Harold Bekkering. 2016. Optimizing Word Learning via Links to Perceptual and Motoric Experience. *Educational Psychology Review* 28, 3 (2016), 495–522. DOI: http://dx.doi.org/10.1007/s10648-015-9334-2

[9] Spencer D. Kelly, Tara McDevitt, and Megan Esch. 2009. Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes* 24, 2 (2009), 313–334. DOI: http://dx.doi.org/10.1080/01690960802365567

[10] James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15).* ACM, New York, NY, USA, 67–74. DOI: http://dx.doi.org/10.1145/2696454.2696457

[11] Sebastian Leitner. 1972. *So lernt man lernen. Der Weg zum Erfolg.* Herder, Freiburg.

[12] Dan Leyzberg, Samuel Spaulding, Mariya Toneva, and Brian Scassellati. 2012. The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2012).* Curran Associates, Inc.

[13] Manuela Macedonia, Karsten Müller, and Angela D. Friederici. 2011. The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping* 32, 6 (2011), 982–998. DOI: http://dx.doi.org/10.1002/hbm.21084

[14] David McNeill. 1985. So you think gestures are nonverbal? *Psychological Review* 92, 3 (1985), 350–371. DOI: http://dx.doi.org/10.1037/0033-295x.92.3.350

[15] Meredith L. Rowe, Rebecca D. Silverman, and Bridget E. Mullan. 2013. The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemporary Educational Psychology* 38, 2 (2013), 109 – 117. DOI: http://dx.doi.org/10.1016/j.cedpsych.2012.12.001

[16] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. Generation and Evaluation of Communicative Robot Gesture. *International Journal of Social Robotics* 4, 2 (2012), 201–217. DOI: http://dx.doi.org/10.1007/s12369-011-0124-9

[17] Maha Salem, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2011. A Friendly Gesture: Investigating the Effect of Multimodal Robot Behavior in Human-Robot Interaction. In *2011 RO-MAN.* Institute of Electrical and Electronics Engineers (IEEE). DOI: http://dx.doi.org/10.1109/roman.2011.6005285

[18] Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. 2017. Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making. In *Proceedings of HRI 2017.*

[19] Marion Tellier. 2008. The effect of gestures on second language memorisation by young children. *Gestures in Language Development* 8, 2 (2008), 219–235. DOI: http://dx.doi.org/10.1075/gest.8.2.06tel

[20] Elisabeth T. van Dijk, Elena Torta, and Raymond H. Cuijpers. 2013. Effects of Eye Contact and Iconic Gestures on Message Retention in Human-Robot Interaction. *International Journal of Social Robotics* 5, 4 (2013), 491–501. DOI: http://dx.doi.org/10.1007/s12369-013-0214-y

[21] Paul Vogt, Mirjam de Haas, Chiara de Jong, Peta Baxter, and Emiel Krahmer. in press. Child-Robot Interactions for Second Language Tutoring to Preschool Children. *Frontiers in Human Neuroscience* (in press).

[22] Lev Vygotsky. 1978. *Mind in society: The development of higher psychological processes.* Harvard University Press, Cambridge, MA.

[23] Rolf A. Zwaan, Robert A. Stanfield, and Richard H. Yaxley. 2002. Language Comprehenders Mentally Represent the Shapes of Objects. *Psychological Science* 13, 2 (2002), 168–171. DOI: http://dx.doi.org/10.1111/1467-9280.00430

[1] http://l2tor.eu

Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations

James Kennedy* Plymouth University, U.K.
Séverin Lemaignan Plymouth University, U.K.
Caroline Montassier INSA Rouen, France
Pauline Lavalade Université Pierre et Marie Curie, France
Bahar Irfan Plymouth University, U.K.
Fotios Papadopoulos Plymouth University, U.K.
Emmanuel Senft Plymouth University, U.K.
Tony Belpaeme Plymouth University, U.K. Ghent University, Belgium

ABSTRACT

An increasing number of human-robot interaction (HRI) studies are now taking place in applied settings with children. These interactions often hinge on verbal interaction to effectively achieve their goals. Great advances have been made in adult speech recognition and it is often assumed that these advances will carry over to the HRI domain and to interactions with children. In this paper, we evaluate a number of automatic speech recognition (ASR) engines under a variety of conditions, inspired by real-world social HRI conditions. Using the data collected we demonstrate that there is still much work to be done in ASR for child speech, with interactions relying solely on this modality still out of reach. However, we also make recommendations for child-robot interaction design in order to maximise the capability that does currently exist.

Keywords

Child-Robot Interaction; Automatic Speech Recognition; Verbal Interaction; Interaction Design Recommendations

1. INTRODUCTION

Child-robot interaction is moving out of lab and into 'the wild', contributing to domains such as health-care [2], education [15, 25], and entertainment [20]. An increasing amount is being understood about how to design interactions from a nonverbal behaviour perspective [13, 14], but many of these domains hinge on effective verbal communication. This includes not only appropriate speech production by robots, but transcribing and understanding speech from young users as well. A prerequisite to this interpretation of speech is having a sufficiently accurate transcription of what is being said.

For this reason, high-quality Automatic Speech Recognition (ASR) is a vital component for producing autonomous human-robot interaction. ASR engines have undergone significant improvements in recent years, particularly following the introduction of new techniques such as deep learning [26]. However, these engines are commonly evaluated against standardised datasets of adult speech [23]. One might naively assume that these improvements will also translate to child speech, and will cope relatively well with noisy (i.e., real-world) environments, such as those experienced in applied HRI. However, this is often observed to not be the case, cf. [19].

In this paper we seek to evaluate the state-of-the-art in speech recognition for child speech, and to test ASR engines in settings inspired by real-world child-robot interactions. We record a variety of pre-determined phrases and spontaneous speech from a number of children speaking English using multiple microphones. We separate recordings by whether they are comparatively clean, or contain noise from the real-world environment. Through consideration of the results, we highlight the limitations of ASR for child speech, and also make a number of interaction design recommendations to maximise the efficacy of the technology currently available.

2. BACKGROUND

Speech recognition has undergone significant advances, building on or moving on from the use of Hidden Markov Models (HMM) towards using deep neural networks (DNN). DNNs have been shown to outperform older HMM based approaches by some margin against standard benchmarks [12]. For example, in a Google speech recognition task a deep neural network reduced the Word Error Rate (WER) to 12.3%, a 23% relative improvement on the previous state-of-the-art [12].

*Corresponding author: james.kennedy@plymouth.ac.uk

HRI '17, March 06 - 09, 2017, Vienna, Austria

ISBN 978-1-4503-4336-7/17/03…$15.00

DOI: http://dx.doi.org/10.1145/2909824.3020229

However, these benchmarks are based on adult speech corpora, such as the TIMIT corpus [17]. It has been noted by other researchers that there is a lack of corpora for children's speech, leading to a lack of training data and a lack of benchmarking for children's speech recognition models [5, 9, 11]. It is commonly assumed that the recent improvements observed in adult speech recognition mean that child speech recognition improved at the same pace, and recognising children's utterances can be achieved with a similar degree of success. However, anecdotal evidence suggests that this is not the case; Lehman et al. [19] state that recognition of children's speech "remains an unsolved problem", calling for research to be undertaken to understand more about the limitations of ASR for children to ease interaction design.

Children's speech is fundamentally different from adult speech: the most marked difference being the higher pitched voice, due to children having a shorter, immature vocal tract. In addition, spontaneous child speech is marked by a higher number of disfluencies and, especially in younger children, language utterances are often ungrammatical (e.g., "The boy *putted* the frog in the box"). As such, typical ASR engines, which are trained on adult speech, struggle to correctly recognise children's speech [8, 24]. An added complexity is caused by the ongoing development of the vocal apparatus and language performance in children: an ASR engine trained for one age group is unlikely to perform well for another age group.

There have been various attempts to remedy this, from adapting adult-trained ASR engines to the spectral characteristics of children's speech [18, 22], to training ASR engines on child speech corpora [6, 8, 10], or combinations of both. For example, Liao et al. [21] have used spoken search instructions from YouTube Kids to train DNNs with some success, resulting in a WER between 10 and 20%. In [24] vocal-tract length normalisation (VTLN) and DNN are used in combination, and when trained on read speech of children aged between 7 and 13 years, result in a WER of approximately 10%. It should be noted that these results are achieved in limited domains, such as spoken search instructions, read speech, or number recognition [22]. Also, the circumstances in which the speech is recorded are typically more controlled than interactions encountered in HRI, where ambient noise, distance and orientation to the microphone, and language use are more variable.

Whilst children's speech recognition in general is a challenge, HRI brings further complexities due to factors such as robot motor noise, robot fan noise, placement and orientation of microphones, and so on. Many researchers adopt interaction approaches that do not rely on verbal interaction due to the unreliability of child ASR, particularly in 'wild' environments. Wizard of Oz (WoZ) approaches have proven popular to substitute for sub-optimal speech recognition and natural language interaction, but when autonomy is important, WoZ is impractical and the use of mediating interfaces to substitute for linguistic interaction has proven successful. Touchscreens, for example, can serve as interaction devices, they provide a focus for the interaction while constraining the unfolding interaction [1]. However, if we wish the field to continue to progress into real-world environments, then it is unrealistic to exclude verbal interaction due to the prevalence of this communication channel in natural interaction.

3. RESEARCH QUESTIONS

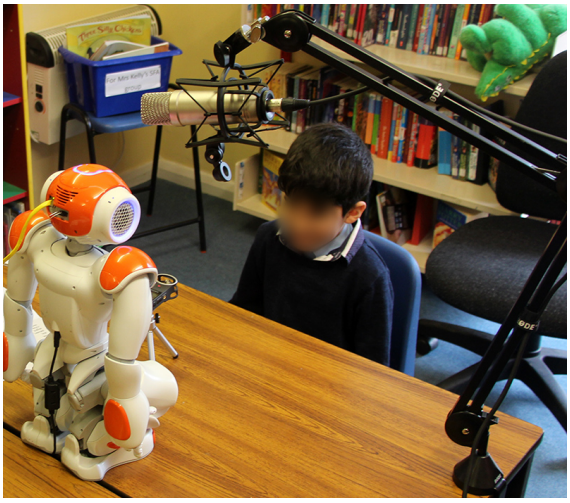


Figure 1: Equipment layout for recording children in a school. The Aldebaran NAO is turned on (but not moving) and records to a USB memory card. The studio microphone and portable microphone record simultaneously.

The previous section highlights that the current performance of ASR for child speech remains unclear. We wish to address this by exploring different variables in the context of child speech, such as the type of microphone, the physical location of the speaker relative to a robot, and the ASR engine. These variables motivate a set of research questions presented below, all in the context of child speech. Their evaluation will be conducted with the aim of producing evidenced guidelines for designing verbal human-robot interactions with children.

Q1 Do external microphones produce better results than robot-mounted microphones?

Q2 How can physical interaction setups be optimised for ASR?

Q3 Is there a benefit to using cloud-based or off-board ASR engines compared to a stock robot ASR engine?

Q4 What is the impact of 'real-world' noise on speech recognition in an HRI inspired scenario?

4. METHODOLOGY

In order to address the research questions posed in the previous section, a data collection and testing procedure was designed. At the time of writing, no corpus of child speech suitable for the intended analysis was publicly available. As such, there is a need for the collection of this data; the procedure for this will be outlined here.

4.1 Participants

A total of 11 children took part in our study, with an average age M=4.9, SD=0.3; 5F/6M. The age group is motivated by the many large-scale initiatives in the US, Europe and Japan exploring linguistic interactions in HRI [2, 3, 19, 20, 25], and the fact that this age group is preliterate, so cannot interact using text interfaces. All children had age-appropriate competency in speaking English at school. All participants

gave consent to take part in the study, with the children's parents providing additional consent for participation, and recording and using the audio data. The children were rewarded after the study with a presentation of social robots.

4.2 Data Collection

In order to collect a variety of speech utterances, three different categories were devised: single word utterances, multi-word utterances, and spontaneous speech. The single word and multi-word utterances were collected by repeating after an experimenter. This was done to prevent any issues with child reading ability. Spontaneous speech was collected through retelling a picture book, 'Frog, Where Are You?' by Mercer Mayer, which is a common stimulus for this activity in language development studies [4]. The single word utterances were numbers from 1 to 10, and the multi-word utterances were based on spatial relationships between two nouns, for example, 'the horse is in the stable'. Five sentences of this style were used; the full set can be downloaded from [16].

The English speech from children was collected at a primary school in the U.K. This served two purposes: firstly, to conduct the collection in an environment in which the children are comfortable, and secondly, to collect data with background noise from a real-world environment commonly used in HRI studies, e.g., [15]. An Aldebaran NAO (hardware version 5.0 running the NaoQi 2.1.4 software) was used as the robotic platform. This was selected as it is a commonly used platform for research with children, as well as for its microphone array and commercial-standard speech recognition engine (provided by Nuance). The robot would record directly from the microphones to a USB memory stick. Simultaneously, a studio grade microphone (Rode NT1-A) and a portable microphone (Zoom H1) were also recording. The studio microphone was placed above the robot and the portable microphone just in front of the robot (Fig. 1).

4.3 Data Processing

Encoding and Segmentation.

All audio files were recorded in lossless WAV format (minimum sampling rate of 44kHz). The audio files from each of the three microphones were synchronised in a single Audacity project. The audio files were then split to extract segments containing the speech under consideration. These segments were exported as lossless WAV files, resulting in 16 files per microphone (48 in total) per child. The spontaneous speech was transcribed and split into sentences. This produced a total of 222 spontaneous speech utterances of various lengths ($M = 7.8$ words per utterance, $SD = 2.6$). The full dataset (audio files and transcripts) is available online at [16].

Noisy vs. Clean Audio Recordings.

As the recordings of children in English were collected during the course of a school day, there is a range of background noise. To study the impact of noise on ASR performance, it is desirable to separate the recordings into those that have minimal background noise ('clean' recordings) and those that have marked background noise ('noisy' recordings). Some noise is unavoidable, or would be present in any HRI scenario, such as robot fan noise, so these were considered 'clean'. Other noise, such as birds outside, other children shouting from the adjacent room, doors closing, or coughing would be considered 'noisy'. This means that the clean recordings

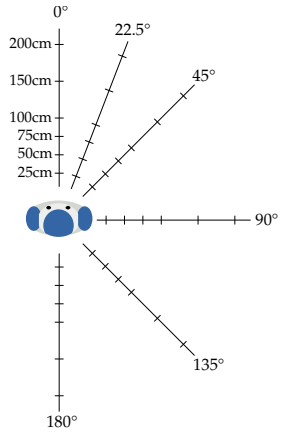


Figure 2: Locations at which speech it played to the NAO to explore how the physical layout of interactions may influence speech recognition rates.

are not noise-free like those from a studio environment, but are a realistic representation of a minimal practical noise level in a 'wild' HRI scenario, thereby allowing us to evaluate recognition accuracy with greater veracity.

To appropriately categorise the recordings as clean or noisy, each one was independently listened to by 3 human coders with the guidance from above as to what is considered clean vs. noisy. Overall agreement levels between coders was good, with Fleiss $\kappa = .74$ (95% CI [.65,.84]) for the fixed utterances and $\kappa = .68$ (95% CI [.60,.75]) for the spontaneous utterances. A recording was categorised as noisy or clean if all 3 coders agreed it was respectively noisy or clean. Where there was any disagreement between coders, the recordings were omitted from analysis of noise impact (59 fixed and 54 spontaneous utterances were excluded). This resulted in 80 noisy recordings, and 37 clean recordings being analysed from the fixed utterances set and 83 clean/85 noisy recordings from the spontaneous utterances set. For some children, the NAO recording failed due to technical difficulties. Therefore, when comparing across microphones, the fixed utterance selection is reduced to 29 clean recordings and 60 noisy recordings.

Manipulation of the Sound Location.

To evaluate the impact of distance and angle on speech recognition, it was necessary to vary the distance between the robot and child, while at the same time keeping the speech utterances constant. As children struggle to exactly reproduce speech acts and over 500 utterances are needed to be recognised, we used pre-recorded speech played through an audio reference speaker (the PreSonus Eris E5) placed at different locations around the robot. In order to match the original volume levels, a calibration process was used where a recording would be played and re-recorded at the original distance between the child and the robot. The audio signal amplitudes between the original and recorded file were then compared. The speaker volume was iteratively revised until the amplitudes matched. This volume was then maintained as the speaker was moved to different distances and angles from the robot, while always facing the robot (to address, at least in part, Q2 from Sec. 3); see Fig. 2 for a diagram of these positions.

4.4 Measures

For recognition cases where a *multiple choice* grammar is used (i.e., the list of possible utterances is entirely predefined, and the recognition engine's task is to pick the correct one), the recognition percentage is used as the metric. Each word or sentence correctly recognised adds 1; the final sum is divided by the number of tested words or sentences. All Confidence Intervals calculated for the recognition percentage include continuity correction using the Wilson procedure. We use the same metric when using template-based grammars (Sec. 5.2.1).

For the cases in which an open grammar is used, we use the Levenshtein distance as a metric at the *letter* level. This decision was made as it reduces punishment for small errors in recognition, which would typically not be of concern for HRI scenarios. For example, when using the Levenshtein distance at the word level (as with Word Error Rate), if the word 'robots' is returned for an input utterance of 'robot', this would be scored as completely unrecognised. At the letter level, this would score a Levenshtein distance of 1, as only a single letter needs to be inserted, deleted or substituted (in this example, the letter 's') to get the correct result. To compare between utterances, normalisation by the number of letters in the utterance is then required to compensate for longer inputs incurring greater possibility of higher Levenshtein distances.

5. RESULTS

This section will break down the results and analysis such that the research questions are addressed. The results are split into two main subsections concerning: 1) technical implementation details, and 2) general ASR performance. The intention is to then provide a practical guide for getting the best performance from ASR in HRI scenarios, as well as an indication of the performance level that can be expected more generally for child speech under different circumstances.

5.1 Technical Best Practices

Throughout this subsection, the ASR engine will remain constant so that other variables can be explored. In this case, the ASR engine used is the one that comes as default on the Aldebaran NAO, provided by Nuance (VoCon 4.7). A grammar is provided to this engine, consisting of numbers (as described in Sec. 4.2) and single word utterances. Longer utterances, along with open grammar and spontaneous speech will be explored in the subsequent subsection.

5.1.1 Type of microphone

Upon observation of the results it became clear that the robot-mounted microphone was vastly outperforming the portable and studio microphones. When visually comparing the waveforms, there was a noticeable difference in recorded amplitude between the NAO signal and the other two microphones. This was despite the standalone microphone input gains being adjusted to maximise the signal (whilst preventing peak clipping). To increase the signal amplitude whilst maintaining the signal-to-noise ratio, the files were normalised. This normalisation step made a significant difference to the results of the speech recognition. For the portable microphone, the recognition percentage after normalisation (70%, 95% CI [59%,79%]) was significantly improved compared to before normalisation (2%, 95% CI [0%,9%]);

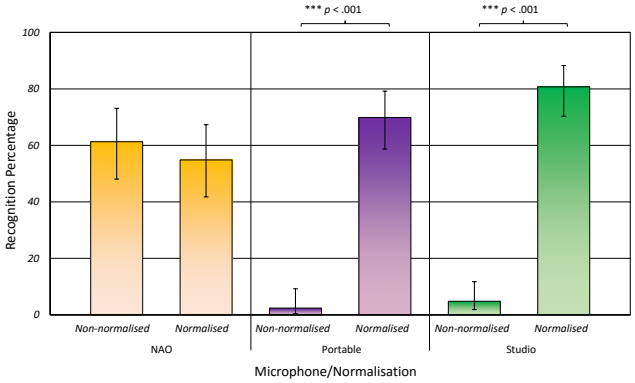


Figure 3: A comparison of recognition percentage of English words and short sentences spoken by children, split by microphone before and after normalisation. * indicates significance at the $p<.001$ level. The recognition is much improved for the portable and studio microphones following normalisation.**

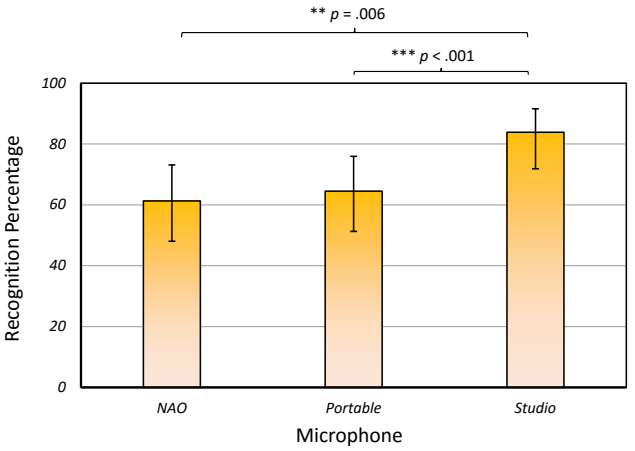


Figure 4: Recognition percentage of numbers spoken by children, split by microphone type (62 utterances). * indicates significance at the $p<.001$ level, ** indicates significance at the $p<.01$ level. The studio microphone provides the best ASR performance, but the difference between on- and lower quality offboard microphones is relatively small.**

Wilcoxon signed-rank test[1] $Z = -7.483, p < .001, r = 0.67$. A similar improvement was observed for the studio microphone when comparing before (5%, 95% CI [2%,12%]) and after (81%, 95% CI [70%,88%]) normalisation; $Z = -7.937, p < .001, r = 0.71$ (Fig. 3). This suggests that the NAO microphones are tuned to maximise the speech level, and if external microphones are to be used, then normalisation of the recordings should be considered a vital step in processing prior to sending to an ASR engine. Therefore, for the remainder of the analysis here, only normalised files are used for the studio and portable microphones.

In exploring Q1, it is observed that the differences between

[1]Due to the recognition being binary on single word inputs, the resulting distributions are non-normal, so non-parametric tests are used for significance testing.

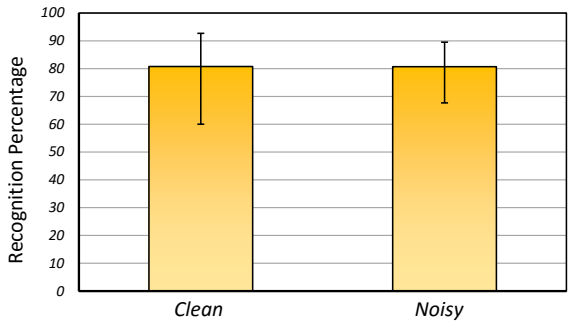


Figure 5: Recognition percentage of single word utterances spoken by children, split by background noise level (83 total utterances). Noise level does not have a significant effect on the recognition rate.

microphones is smaller than may have been expected. The NAO microphones are mounted in the head of the robot near a cooling fan which produces a large amount of background noise. It could therefore be hypothesised that the ASR performance would greatly increase by using an off-board microphone, and that using a higher-quality microphone would improve this further. Using Friedman's test, a significant difference at the $p < .05$ level is found between the NAO (61%, 95% CI [48%,73%]), portable (65%, 95% CI [51%,76%]), and studio (84%, 95% CI [72%,92%]) microphones; $\chi^2(2) = 9.829, p = .007$. Post-hoc Wilcoxon signed-rank pairwise comparisons with Bonferroni correction reveal a statistically significant difference between the portable and studio microphones ($Z = -3.207, p < .001, r = 0.29$; Fig. 4), and between the NAO and studio microphones ($Z = -2.746, p = .006, r = 0.24$). Differences between the portable and NAO microphones ($Z = -0.365, p = .715, r = 0.03$) were not significant. This suggests that there is no intrinsic value to using an off-board microphone, but that a high quality off-board microphone can improve the ASR results. The difference between the robot microphone and the external studio grade microphone is fairly substantial, with a recognition percentage improvement of around 20%point ($r = 0.28$). It would be scenario specific as to whether the additional technical complexity of using a high-quality external microphone would be worth this gain, and indeed, in scenarios where the robot is mobile, use of a studio grade microphone may not be a practicable option.

5.1.2 *Clean vs. Noisy Recording Environment*

Splitting the files by whether they were judged to be clean or noisy (as described in Sec. 4.3), it was observed that the noise did not appear to have a significant impact on the results of the ASR. Using the studio microphone (i.e., the best performing microphone) for the number utterances, a Mann-Whitney U test reveals no significant difference between clean (81%, 95% CI [60%,93%]) and noisy (81%, 95% CI [68%,90%]) speech; $U = 740.5, p = .994, r = 0.00$ (Fig. 5). The apparent robustness of the ASR engine to noise is of particular benefit to HRI researchers given the increasingly 'real-world' application of robots, where background noise is often near impossible (nor desirable) to prevent.

However, this does not mean that noise does not play a role in recognition rates. In this instance, the ASR engine is restricted in its grammar; the effect of noise in open grammar

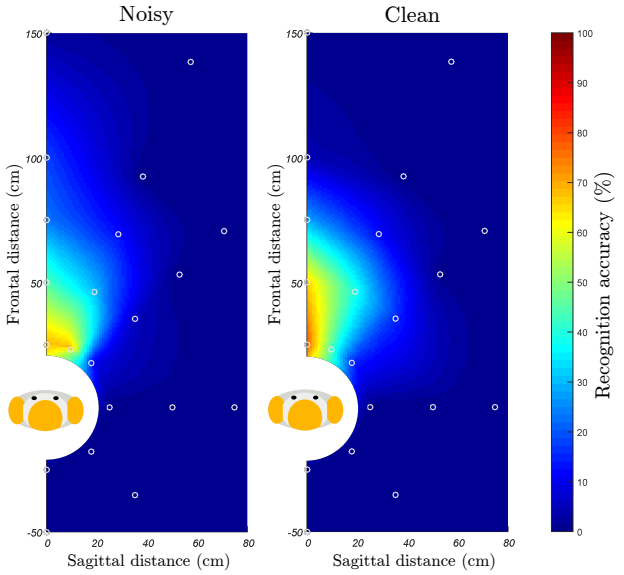


Figure 6: Interpolated heatmap of recognition percentage as a function of distance and orientation to the robot. Interpolation has been performed based on the measurements made at the small white circles. On the *left* is the heatmap for the noisy audio, whereas the *right* is for clean audio. The clean audio is better recognised at further distances from the robot, however, in both cases, recognition accuracy is 0% to the side and behind the robot.

situations is explored in the next subsection. Additionally, when the distance of the sound source to the microphone is varied, background noise becomes a greater factor.

5.1.3 *Sound Source Location*

Measurements were made as in Fig. 2 using the built in NAO microphone, with the replayed audio from the studio microphone (as described in Sec. 4.3). Due to the number of data points this generates (540 per child), the findings in full will not be produced here, but to get a high-level picture of how the distance and orientation influences recognition rates, a heatmap can be seen in Fig. 6.

Two observations can be made from this data that have particular relevance for HRI researchers. The first is the platform-specific observation that with the NAO robot (currently one of the most widely used research platforms for social HRI) the utterance recognition rate drops dramatically once the sound source reaches a 45 degree angle to the

Distance (cm)	Clean % [95% CI]	Noisy % [95% CI]
25	73 [52,88]	77 [64,87]
50	65 [44,82]	44 [31,58]
75	27 [12,48]	23 [13,36]
100	4 [0,22]	18 [9,30]

Table 1: ASR recognition rates for children counting from one to ten. Recordings were played frontally at different distances from the robot. Note how recognition falls sharply with distance when the speech contains noise.

robot head, and becomes 0 once it reaches 90 degrees. The implication of this is that when using the NAO, it is vital to rotate the head to look at the sound source in order to have the possibility of recognising the speech. This is of course dependent on the current default software implementation; four channels of audio exist, but for ASR only the front two are used, and so a workaround could be created for this. The second, broader observation, is that the background noise and distance seem to influence recognition rates when combined. Fig. 5 shows how little impact noise has when the files are fed directly into the robot ASR, but when combined with distance, there is a marked difference beyond 50cm. Table 1 shows the measurements for the first metre directly in front of the robot; at 25cm the difference between clean and noisy files are minimal, however at 50cm, the difference is more pronounced, with recognition rates dropping fast.

5.2 ASR Performance with Children

The previous subsection addressed variables in achieving a maximal possible speech recognition percentage through modifying the technical implementation, such as different microphones, distances to a robot, orientation to a robot, and background noise levels. This subsection will provide a complementary focus on exploring the current expected performance of ASR with children under different speech and ASR engine conditions. This will include a comparison of differing length utterances, spontaneous utterances, and different ASR engines with varying grammar specifications. For all analyses in this section, the studio microphone signal is used to provide the best quality sound input to the speech engines (and provide a theoretical maximal performance).

5.2.1 Impact of Providing a Grammar

Tests on child speech in the previous subsection were performed with single word utterances, with a grammar consisting of only those utterances. This kind of multiple choice is relatively straightforward, and this carries over to slightly longer utterances too. We compare the recognition rate of the fixed multi-word utterances (34 spatial relation sentences as described in Sec. 4.2) under 3 conditions using the built-in NAO ASR: 1) with a fixed grammar containing the complete utterances, e.g., "one" or "the dog is on the shed" (i.e., multiple choice), 2) with a template grammar for the sentences (as seen in Fig. 8), and 3) with an open grammar. This progressively reduces the prior knowledge the ASR engine has about what utterances to expect. The full mix of noisy and clean utterances were used as there was no observed significant correlation in any of the three conditions between ASR confidence level and noise condition, nor between noise condition and resulting recognition rates. The grammar condition has a significant impact on the recognition percentage; Friedman's test $\chi^2(2) = 39.92, p < .001$. Post-hoc Wilcoxon signed-rank pairwise comparisons with Bonferroni correction reveal a statistically significant difference between the multiple choice (74%, 95% CI [55%,86%]) and template grammars (53%, 95% CI [35%,70%]); $Z = -2.646, p = .008, r = 0.32$. The template grammar in turn offers a significant improvement over the open grammar (0%, 95% [0%,13%]); $Z = -4.243, p < .001, r = 0.51$ (Fig. 7).

5.2.2 Comparison of ASR Engines

Finally, we look at how different ASR engines perform, under identical recording conditions. We compare the Google

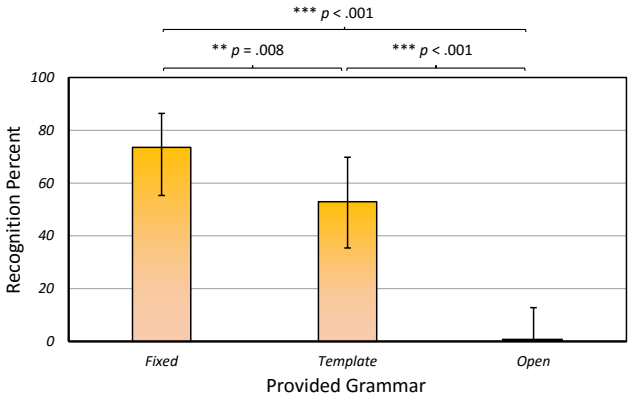


Figure 7: Recognition percentage when providing a fixed grammar, a template grammar, and an open grammar on short utterances. The fixed 'multiple choice' grammar produces the best recognition, followed by a template. The open grammar, on average, recognises almost no sentences correctly.

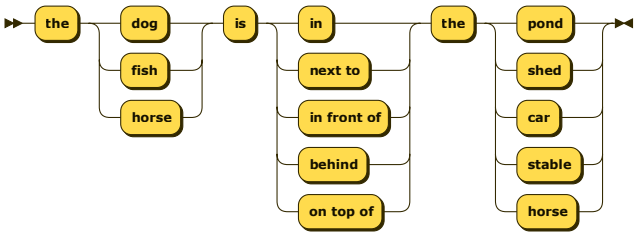


Figure 8: Template for the grammar provided to the ASR for the fixed utterances. 75 different sentences can be generated from this grammar.

Speech API (as found in the Chrome web browser for instance), the Microsoft Speech API (as found in the Bing search engine), CMU PocketSphinx, and the NAO-embedded Nuance VoCon 4.7 engine; studies were run in August 2016. The audio samples are those recorded with the studio microphone; they include native and non-native speakers as well as noisy and clean samples; they include both the fixed sentences and the spontaneous speech; no grammar is provided to the engine (i.e., open grammar).

As performing recognition with an open grammar is a much harder challenge for recognition engines, the recognition percentage alone is no longer a sufficient measurement to compare between performance of ASR engines due to the very low number of exact utterance recognitions across all engines. Instead we use the Levenshtein distance (LD) at the letter level. As the utterance length for the spontaneous speech is also variable, the Levenshtein distance is normalised by utterance length (as per Sec. 4.4). This provides a value between 0 and 1, where 0 means the returned transcription matches the actual utterance, and 1 means not a single letter was correct. Values in between indicate the proportion of letters that would have to be changed to get the correct response, therefore lower scores are better. Table 2 provides one recognition example with the corresponding Levenshtein distances.

While the LD provides a good indication of how close the result is from the input utterance, the examples in Table 2

Google API	*then the wraps looks at the dog*	[LD=0.17]
Microsoft API	*rat look at dogs*	[LD=0.48]
PocketSphinx	*look i personally*	[LD=0.83]

Table 2: Recognition results and Levenshtein distance for three ASR engines on the input utterance "then the rat looked at the dog". The NAO-embedded Nuance engine did not return any result.

evidence that this metric does not necessarily reflect *semantic* closeness. In this particular case, the Bing result "rat look at dogs" is semantically closer to the original utterance than the other answers. For this reason, we assess recognition performance in open grammar using a combination of three metrics: 1) the Levenshtein distance; 2) raw accuracy (i.e., the number of exact matches between the original utterance and the ASR result); 3) a manually-assessed 'relaxed' accuracy. The utterance would be considered accurate in the 'relaxed' category if small grammatical errors are present, but not semantic errors. Grammatical errors can include pluralisation, removal of repetitions, or small article changes ('the' instead of 'a'). For example, if an input utterance of "and then he found the dog" returned the result "and then he found a dog", this would be considered accurate, however "and then he found the frog" would produce a similar LD, but the semantics have changed, so this would not be included in the relaxed accuracy category.

Table 3 shows that when the input utterance set is changed to use spontaneous speech, the average normalised LD does not change much for any of the ASR engines. Nor do the LD rates change much when only clean spontaneous speech is used, providing further evidence for the minimal impact of noise as established in Sec. 5.1.2. However, there is a marked difference between Google and the other recognition engines. The average LD from Google is around half that of the other engines, and the number of recognised sentences in both the strict and relaxed categories is substantially higher. The recognition performance remains however generally low: using relaxed rules, the currently best performing ASR engine (Google Speech API) for our data recognises only about 18% of a corpus of 222 child utterances (utterances have a mean length of $M = 7.8$ words, $SD = 2.6$).

To help decide whether or not the results returned from Google would actually be usable in autonomous HRI scenarios, it is necessary to determine when the utterance is correctly recognised. This is typically indicated through the *confidence value* returned by the recognition engine. To further explore this, we assess the number recognition percentage at different thresholds within the confidence level (Fig. 9). A total of 101 results from the 222 passed to the recogniser returned a confidence level (a confidence value is not returned when the uncertainty of the ASR engine is too high). To achieve just below 50% semantically correct recognition accuracy, the confidence threshold could be set to 0.8, which would only include 36 utterances. While a clear improvement over the 18% previously achieved when not taking into consideration the confidence value, a 50% recognition rate is arguably not sufficient for a smooth child-robot verbal interaction, and would still require the system to reject nearly 2/3 of the child utterances.

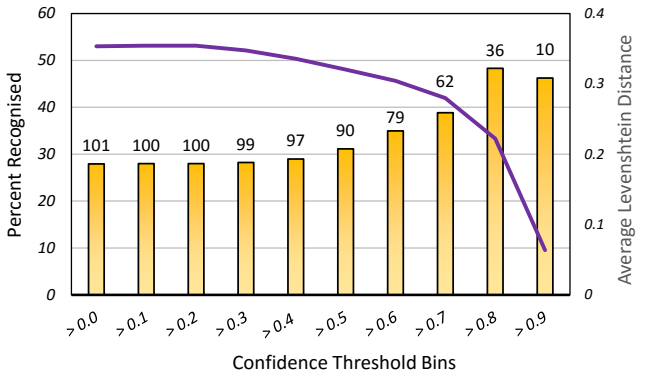


Figure 9: Histogram of recognition percentage (using the relaxed, manually coded criteria) for spontaneous speech grouped by confidence levels (indicated by the number above each bar) returned by Google ASR. The average Levenshtein Distance is also shown on the secondary axis. Recognition increases with higher confidence ranges, but few utterances have a high confidence.

6. DISCUSSION

Our results show that, at the time of writing, automatic speech recognition still does not work reliably with children, and should not be relied upon for autonomous child-robot interaction.

Speech segmentation is one aspect that we did not investigate. The segmentation of speech units and rejecting non-speech parts is an important factor in speech recognition. For example, noise can be mistakenly recognised by ASR engines as speech, or a pause in the middle of a sentence might interrupt the segmentation. Existing solutions (like a beep sound indicating when to talk) are not ideal for children of this age. Our manual segmentation likely leads to better results than would be expected with automatic segmentation.

We did not analyse if gender had an effect on ASR due to the age of the children used in the study. It has been shown that there are no significant differences in the vocal tract between genders at the age under consideration (5-6 years old) [7], so we do not expect differing performance based on gender.

Mitigation strategies for poor ASR performance depend on the ASR engine. We have specifically investigated the use of constrained grammar with the NAO's Nuance engine; and the use of the recognition confidence with the Google ASR. While severely constraining the interaction scope, none of these techniques were found to provide satisfactory results. In our most favourable test case (children speaking numbers from one to ten in front of the robot, at about 25cm; the robot having an explicit 'multiple choice' grammar), the ASR would return an incorrect result in one of four cases, and could not provide any meaningful confidence value. This result is disappointing, particularly when considering that interactions based on 'multiple choice' are difficult to rely on with children, as they tend not to remember and/or comply to the given set of recognisable utterances.

Template-based grammars (or 'slot-filling' grammars) where the general structure of the sentence is known beforehand, and only a limited set of options are available to fill the 'gaps' are a potentially interesting middle-ground between

	Google		Bing		Sphinx		Nuance	
	M LD [95%CI]	*% rec.*	*M* LD [95%CI]	*% rec.*	*M* LD [95%CI]	*% rec.*	*M* LD [95%CI]	*% rec.*
fixed (n=34)	**0.34** [0.24,0.44]	*11.8 [38]*	0.64 [0.56,0.71]	*0 [0]*	0.68 [0.64,0.73]	*0 [0]*	0.76 [0.73,0.80]	*0 [0]*
spontaneous (n=222)	**0.39** [0.36,0.43]	*6.8 [17.6]*	0.64 [0.61,0.67]	*0.5 [2.4]*	0.80 [0.77,0.84]	*0 [0]*	0.80 [0.78,0.82]	*0 [0]*
spontaneous clean only (n=83)	**0.40** [0.35,0.45]	*6.0 [16.9]*	0.63 [0.58,0.68]	*1.2 [1.2]*	0.78 [0.72,0.85]	*0 [0]*	0.78 [0.75,0.81]	*0 [0]*

Table 3: Comparison between four ASR engines using fixed, all spontaneous, and clean spontaneous speech utterances as input. Mean average normalised Levenshtein Distance (*M* LD) indicates how good the transcription is. *% rec* indicates the percentage of results that are an exact match for the original utterance, with the values in square brackets *[ ]* indicating matches with 'relaxed' accuracy.

'multiple choice' grammars and open speech. However, we show that in our test case (grammar depicted in Fig. 8), the correct utterance was recognised in only 50% of the cases, again without any useful confidence value.

In the realm of open grammars, the Google Speech API returned the most accurate results by a large margin. When run on grammatically correct, regular sentences (the ones generated from the grammar depicted in Fig. 8), it reaches 38% accuracy in recognition when minor grammatical differences are allowed. This result, while likely not yet usable in today's applications, is promising. However, when looking at children's spontaneous speech, the recognition rate drops sharply (to around 18% of successful recognition). This difference can be explained by the numerous disfluencies and grammatical errors found in natural child speech. To provide an example, a relatively typical utterance from our data was "and... and the frog didn't went to sleep". The utterance has a repetition and disfluency at the start, and is followed by grammatically incorrect content. This is, in our opinion, the real challenge that automatic child speech recognition faces: the need to account for the child-specific language issues, beyond the mere differences between the acoustic models of adults vs. children. This is a challenge not only for speech-to-text, but as well for later stages of the verbal interaction, like speech understanding and dialogue management.

Our results allow us to make a number of recommendations for designing child-robot interaction scenarios that include verbal interaction. Most of these are also applicable to adult settings and would be expected to contribute to a smoother interaction.

- Constrain the interaction by leading the child to a limited set of responses. This typically works well for older children, but carries the risk of making the interaction stale.
- Use additional input/output devices. A touchscreen has been found to be a particularly effective substitute for linguistic input [1, 14], but also other devices –such as haptic devices– should be considered.
- Place the young user in the optimal location for ASR. The location and orientation relative to the microphone (and robot) has a profound impact on ASR performance (Sec. 5.1.3). A cushion, stool or chair can help children sit in the optimal location.
- Constrain the grammar of the ASR. While not all ASR engines allow for this (cf. Bing), some will allow constraints or "hints" on what is recognised. This proves to be valuable in constrained interaction settings, for example, when listening only for numbers between 1 and 10 (Sec. 5.2.1).
- Background noise appears to be less of an issue than initially anticipated. It appears that the current ASR engines have effective noise cancelling mechanisms in place. Nevertheless, "the less noise, the better" remains true, particularly when interacting at a distance from the robot (Sec's 5.1.2 & 5.1.3).
- A lack of ASR performance does not mean that the robot should not produce speech, as speech has been found to be particularly effective to engage children.

We opted to evaluate the ASR capabilities of the Aldebaran NAO platform, as it is the most commonly used robot in commercial and academic HRI. While the NAO system under performs for child speech, some performance could be gained through using a high-quality external microphone and cloud-based ASR, with Google as clear favourite.

7. CONCLUSION

Language is perhaps the most important modality in human-to-human interaction and as such, functional natural language interaction forms a formidable prize in human-machine interaction. Speech recognition is the entry point to this and while there has been steady progress in speaker-independent adult speech recognition, the same progress is currently lacking from children's speech recognition. For various reasons –pitch characteristics of children's voices, speech disfluencies, and unsteady developmental changes– child speech recognition is expected to require a multi-pronged approach and recognition performance in unconstrained domains is currently too low to be practical.

This has a profound impact on the interaction between children and technology, especially where pre-literacy children are concerned, typically ages 6 and younger. As they have no means of entering input other than by speaking to the device, the interaction with pre-literacy children stands or falls with good speech recognition.

Our results show that natural language interactions with children are not yet practicable. Today, building rich and natural interactions between robots and children still requires a complex alchemy: a careful design of the interaction that

leads the responses of the young user in such a way that restrictive ASR grammars are acceptable, the understanding and production of rich non-verbal communication cues like gaze, and a judicious use of supporting technology such as touchscreens.

8. ACKNOWLEDGEMENTS

This work has been supported by the EU H2020 L2TOR project (grant 688014), the EU FP7 DREAM project (grant 611391), and the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227).

9. REFERENCES

[1] P. Baxter, R. Wood, and T. Belpaeme. A touchscreen-based 'sandtray' to facilitate, mediate and contextualise human-robot social interaction. In *Proceedings of the 7th annual ACM/IEEE international conference on Human-Robot Interaction*, pages 105–106. ACM, 2012.

[2] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, R. Looije, M. Neerincx, Y. Demiris, R. Ros-Espinoza, A. Beck, L. Cañamero, A. Hiolle, M. Lewis, I. Baroni, M. Nalin, P. Cosi, G. Paci, F. Tesser, G. Sommavilla, and R. Humbert. Multimodal Child-Robot Interaction: Building Social Bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

[3] T. Belpaeme, J. Kennedy, P. Baxter, P. Vogt, E. E. Krahmer, S. Kopp, K. Bergmann, P. Leseman, A. C. Küntay, T. Göksun, A. K. Pandey, R. Gelin, P. Koudelkova, and T. Deblieck. L2TOR - second language tutoring using social robots. In *Proceedings of the 1st International Workshop on Educational Robots*, Paris, France, 2015.

[4] R. A. Berman and D. I. Slobin. *Relating events in narrative: A crosslinguistic developmental study.* Psychology Press, 2013.

[5] P. Cosi, M. Nicolao, G. Paci, G. Sommavilla, and F. Tesser. Comparing open source ASR toolkits on Italian children speech. In *Proceedings of the Workshop on Child Computer Interaction*, 2014.

[6] S. Fernando, R. K. Moore, D. Cameron, E. C. Collins, A. Millings, A. J. Sharkey, and T. J. Prescott. Automatic recognition of child speech for robotic applications in noisy environments. *arXiv preprint*, arXiv:1611.02695, 2016.

[7] W. T. Fitch and J. Giedd. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106(3):1511–1522, 1999.

[8] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos. A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, pages 7:1–7:8. ACM, 2009.

[9] P. Grill and J. Tučková. Speech databases of typical children and children with SLI. *PloS one*, 11(3):e0150365, 2016.

[10] A. Hagen, B. Pellom, and R. Cole. Children's speech recognition with application to interactive books and tutors. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2003*, pages 186–191. IEEE, 2003.

[11] A. Hämäläinen, S. Candeias, H. Cho, H. Meinedo, A. Abad, T. Pellegrini, M. Tjalve, I. Trancoso, and M. Sales Dias. Correlating ASR errors with developmental changes in speech production: A study of 3-10-year-old European Portuguese children's speech. In *Proceedings of the Workshop on Child Computer Interaction*, 2014.

[12] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[13] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology*, 6, 2015.

[14] J. Kennedy, P. Baxter, and T. Belpaeme. Nonverbal Immediacy as a Characterisation of Social Behaviour for Human-Robot Interaction. *International Journal of Social Robotics*, in press.

[15] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme. Social Robot Tutoring for Child Second Language Learning. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, pages 67–74. ACM, 2016.

[16] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme. Children speech recording (English, spontaneous speech + pre-defined sentences). Data set, 2016. http://doi.org/10.5281/zenodo.200495.

[17] L. F. Lamel, R. H. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Speech Input/Output Assessment and Speech Databases*, 1989.

[18] L. Lee and R. Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing*, 6(1):49–60, Jan 1998.

[19] J. F. Lehman. Robo fashion world: a multimodal corpus of multi-child human-computer interaction. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 15–20. ACM, 2014.

[20] I. Leite, H. Hajishirzi, S. Andrist, and J. Lehman. Managing chaos: models of turn-taking in character-multichild interactions. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, pages 43–50. ACM, 2013.

[21] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani. Large vocabulary automatic speech recognition for children. In *Proceedings of Interspeech*, 2015.

[22] A. Potamianos and S. Narayanan. Robust recognition of children's speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616, 2003.

[23] M. L. Seltzer, D. Yu, and Y. Wang. An investigation of deep neural networks for noise robust speech recognition. In *Proccedings of the IEEE International*

Conference on Acoustics, Speech and Signal Processing, pages 7398–7402. IEEE, 2013.

[24] R. Serizel and D. Giuliani. Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. *Natural Language Engineering*, 1:1–26, 2016.

[25] F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, and K. Hayashi. Pepper learns together with children: Development of an educational application. In *Proceedings of the IEEE-RAS 15th International Conference on Humanoid Robots*, HUMANOIDS 2015, pages 270–275. IEEE, 2015.

[26] D. Yu and L. Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2015.

Exploring Different Types of Feedback in Preschooler and Robot Interaction

Mirjam de Haas, Peta Baxter, Chiara de Jong, Emiel Krahmer, Paul Vogt
Tilburg center for Cognition and Communication, Tilburg School of Humanities, Tilburg University
Tilburg, the Netherlands
mirjam.dehaas@tilburguniversity.edu

ABSTRACT
This study considered the feedback of a robot during second language tutoring. Traditionally, robots are programmed to provide feedback as teacher; we propose a robot that acts as a peer to motivate preschoolers during the tutoring. We conducted an experiment with 65 preschoolers (M = 3.6 years) in which the robot varied feedback in three conditions: peer-like (explicit negative), adult-like (explicit positive and implicit negative) and no feedback. The results suggest that feedback did not influence children's engagement (measured via eye-gaze), although children who received peer-like feedback seemed to perform more independently during the learning task (requiring less interventions from the experimenter).

Keywords
Social robots; second language tutoring; child-robot interaction

1. INTRODUCTION
Recently, more attention has been given to robots in education, for example to teach children a second language [1], [2]. In such settings the robot is used as an adult teacher, and the ensuing child-robot interactions are based on interactions between children and their teachers. However, in long-term interactions, children may treat the robot as a peer, not as a teacher [3]. Moreover, peer interactions have been shown to have a positive effect on language development [3]. We therefore develop a tutor robot as a more knowledgeable peer, who can adjust the difficulty of the task, give personalized feedback and provide new information, but can also make mistakes, and allows for learning-by-teaching [2].

One of the questions that arises is how should the robot provide adequate feedback during language tutoring, such that is it both pleasant and effective for learning? Adult caregivers normally praise children to encourage them and recast utterances to provide corrective feedback implicitly, but peers may also use explicit negative feedback [5]. Research has shown that explicit negative feedback can have more impact on learning, although positive feedback gives some reassurance to the learner [6].

In this study, our aim is not to investigate the effect of feedback on learning, but instead to investigate how children react to these different types of feedback. We implemented three types of feedback in a NAO-based robot tutor: adult-like feedback, peer-like feedback and no feedback. The adult-like behavior of the robot used reformulations to correct the children ("*Three* means three", the text said in English is indicated in Italics, the rest was said in Dutch) and positive feedback ("Well done!") when they responded correctly. The positive feedback was accompanied by colored eye-LEDs to indicate happiness. The second peer-like condition, only provided explicit negative feedback ("That's wrong!"). In the no feedback condition, the robot did not give any corrections or feedback. We examined how children responded to these different feedback conditions in terms of how engaged they were during the interactions as measured through eye-gaze.

HRI'17 Companion, *March 06–09, 2017, Vienna, Austria.*
ACM 978-1-4503-4885-0/17/03
http://dx.doi.org/10.1145/3029798.3038433

2. EXPERIMENTAL SETTING
We conducted an experiment with 65 three-year old children (30 boys, 35 girls; M = 3.6 years, SD = 0.29) at different preschools in the Netherlands. Six children stopped with the experiment before it was finished and were excluded from the data. The remaining participants were randomly assigned to the three conditions: adult-like feedback (N=20), peer-like feedback (N=19) and no feedback (N=19). In all conditions the experimenter was seated nearby and provided reassurance for the children if necessary. While the experimenter instructed the children to perform a task, or occasionally provided help if required, she was careful not to provide feedback. Figure 1 shows a participant interacting with the robot and the blocks.

Figure 1. Experimental Setup.

In the week before to the experiment, all children took part in a group introduction to familiarize them with the NAO robot. During the actual experiment, children were taught the first four count words in English. The interaction consisted of an introductory phase followed by the tutoring session. In the tutoring session, each target word was repeated only once, so the task was repeated four times. However, the children were already exposed to the target words in the introduction phase. The interaction was in Dutch and only the target words were in English. During the experiment, the robot requested the child to collect a certain number of blocks using an English target word. After the child collected the blocks, the robot provided feedback to the child according to the condition. For example, in the adult feedback condition, a correct answer would invoke a happy expression, together with a positive verbal feedback, while in the other conditions the robot would continue to the next step. In the case of a mistake, the child would receive negative feedback and then could try again. The duration of the experiment was between 10 and 15 minutes.

3. RESULTS

3.1 Eye-gaze

The experiment was recorded in order to analyze the participants' interaction with the robot, and in particular their eye-gaze in reaction to the different feedback. We annotated the gaze towards the robot, human experimenter, to the blocks and elsewhere and conducted a repeated measures ANOVA to explore the differences within the groups. In general, the children looked significantly longer at the blocks and the robot than at the experimenter (see Figure 2).

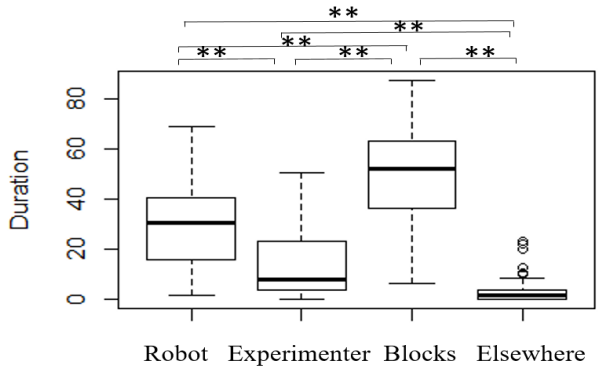


Figure 2. Time spent on gaze direction shown for all conditions combined.

Note. *p<0.05, **p<0.01.

Immediately after the moment that the robot gave instructions, the children looked more often at the robot and the blocks in the adult feedback and the peer feedback conditions, but looked more often at the experimenter in the no feedback condition (see Table 1). Moreover, children received less help from the experimenter in the peer condition than in the adult feedback condition and most in the no feedback condition (22 times, 36 times and 43 times respectively).

However, we did not see any significant differences in the duration of the gaze towards the experimenter, the robot and the blocks across the three conditions.

Table 1. Total number of gaze occurrences towards the experimenter, robot or blocks immediately after the robot gave the instructions

Gaze	Adult feedback	Peer feedback	No feedback
Robot	32	50	35
Experimenter	10	11	50
Blocks	50	38	45

4. DISCUSSION

In this experiment we explored how preschoolers interact with a robot tutor and how they respond to the robot's different types of feedback. Children in the adult feedback condition received most feedback from the robot. Moreover, children in the peer feedback condition received less help from the experimenter, and looked less at the experimenter after receiving the instructions from the robot. According to Spilton and Lee [7] children respond more often to explicit, specific questions than to implicit nonverbal and verbal feedback from peers. This might explain our results, as the children received explicit negative feedback in the peer condition, and required less help from the experimenter.

While the gaze duration results did not show significant differences between the three conditions, the children looked less often at the experimenter in the two feedback conditions. This suggests that children respond well to the robot's feedback. In all conditions, the children looked most at the blocks and the robot. It is possible that the non-significant differences in gaze duration between the conditions are due to individual differences between the children. In general, we saw substantial differences between children in how they responded to the robot, and further exploring these differences is an interesting line for future research.

Importantly, we believe that the implicit and explicit feedback can be useful in a tutoring session, and it would be beneficial for the robot to be able to adapt to the child and the setting with regard to feedback. The implicit negative feedback together with the positive feedback can, for example, be used in cases where the child is demotivated by previous mistakes. The explicit negative feedback may, on the other hand, be used to increase the learning gain of the child.

5. ACKNOWLEDGMENTS

This work has been supported by the EU H2020 L2TOR project (grant 688014). The authors would like to the Kinderopvanggroep Tilburg and all preschools for participating in this research.

6. REFERENCES

[1] Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E.E.J., Kopp, S., Bergmann, K., Leseman, P., Küntay, A.C., Göksun, T., Pandey, A.K., Gelin, R., Koudelkova, P., and Deblieck, T. 2015. L2TOR - Second Language Tutoring using Social Robots. In *Proceedings of the 1st Int. Workshop on Educ. Robots. Springer*

[2] 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 Int. Conf. Human-Robot Interact. HRI '15*.

[3] Mashburn, A.J., Justice, L. M., Downer, J. T., and Pianta, R.C. 2009. Peer effects on children's language achievement during pre-kindergarten. *Child Dev*, 80,3, 686–702.

[4] Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H., 2004. Interactive Robots as Social Partners and Peer Tutors for children: A field trial. *Hum.-Comput Interact.*, 19, 61-84.

[5] Long. M. H. 2006. Recasts in SLA: The story so far," in *Problems in SLA. Second Language Acquisition Research Series.*, Mahwah, NJ: Lawrence Erlbaum Associates. 75-116.

[6] Okita, S.Y., Schwartz, D.L.2013. Learning by Teaching Human Pupils and Teachable Agents: The Importance of Recursive Feedback

[7] Spilton, D., and Lee, L.C. 2016. Some Determinants of Effective Communication in Four-Year-Olds. *Soc. Res. Child Dev. Commun.*, 48,3, 968–977.

Adaptive Robot Language Tutoring Based on Bayesian Knowledge Tracing and Predictive Decision-Making

Thorsten Schodde
CITEC, Bielefeld University
Bielefeld, Germany
tschodde@techfak.uni-bielefeld.de

Kirsten Bergmann
CITEC, Bielefeld University
Bielefeld, Germany
kirsten.bergmann@uni-bielefeld.de

Stefan Kopp
CITEC, Bielefeld University
Bielefeld, Germany
skopp@techfak.uni-bielefeld.de

ABSTRACT

In this paper, we present an approach to adaptive language tutoring in child-robot interaction. The approach is based on a dynamic probabilistic model that represents the interrelations between the learner's skills, her observed behaviour in tutoring interaction, and the tutoring action taken by the system. Being implemented in a robot language tutor, the model enables the robot tutor to trace the learner's knowledge and to decide which skill to teach next and how to address it in a game-like tutoring interaction. Results of an evaluation study are discussed demonstrating how participants in the adaptive tutoring condition successfully learned foreign language words.

CCS Concepts

•Computing methodologies → Probabilistic reasoning; Cognitive robotics; •Applied computing → Interactive learning environments; •Human-centered computing → *Empirical studies in HCI;*

Keywords

Language tutoring; Education; Assistive robotics; Bayesian Knowledge Tracing; Decision making

1. INTRODUCTION

The use of robots for educational purposes has increasingly moved into focus in recent years. This is due to two major developments. First, robots became cheaper and more robust so that applications in everyday environments are now conceivable. In particular, technology has matured up to a point where intuitive interaction using natural language or gesture has become feasible. Second, the need for second language learning becomes increasingly important, and empirical evidence has demonstrated that learning with and from a physically present, interactive robot can be more effective than learning from classical on-screen media [14, 15, 20, 22]. In fact, recent research showed that performance can increase up to 50% [17]. It can, hence, be assumed that tutoring using social robots is qualitatively different from alternative digital tutoring technologies. Nowadays, first practical applications can be found, e.g. in nursery where toy robots teach the alphabet to kids in a very simple way. More generally, findings from a variety of settings seem to suggest that robots can help small children to develop in an educational setting [10, 18, 24, 27].

In the L2TOR project[1], we investigate in how far a social robot can support children at pre-school age with respect to second language learning. Learning a language is a very complex task. It involves not only acquiring vocabulary, but also learning prosodic features, syntactical structures, semantic meanings as well as situation-dependent language use. Yet, it has been argued that social robots can create the interactive environment and motivational experience needed to learn languages [19].

One of the most important aspects in tutoring is the robot's ability to keep track of the knowledge state, i.e. the learned and not-yet-learned skills, of the child interacting with it. This information is indispensable to enable a personalized tutoring interaction and to optimize the learning experience for the child [27]. The tutor has to structure the tutoring interaction, choose the skills to be trained, adjust the difficulty of the learning tasks appropriately and has to adapt its verbal and non-verbal behaviour.

The importance of personalized adjustments in the robot's behaviour has been substantiated in recent research showing that participants who received personalized lessons from a robot (based on heuristic skill assessment) outperformed others who received a non-personalized training [22]. Suboptimal robot behaviour (e.g. too much, too distracting, mismatching or in other ways inappropriate) can even hamper learning [17]. In this paper we present an integrated approach for tracing the knowledge of the learner during a L2 learning interaction together with a strategic adaptation of tutoring actions.

In the following, we discuss related work in Section 2. In Section 3 an extension of Bayesian Knowledge Tracing is presented as well as a model to select the next tutoring actions based on the predicted effects they may have on the learner's knowledge state. This model has been implemented in a robot that provides language tutoring in a game-like fashion. Section 4 introduces the empirical basis for this scenario and observational studies on language tutoring in kindergarten. Section 5 presents an evaluation study carried out with this robot and Section 6 discusses the results.

[1] http://www.l2tor.eu

HRI '17, March 06 - 09, 2017, Vienna, Austria

ISBN 978-1-4503-4336-7/17/03…$15.00

DOI: http://dx.doi.org/10.1145/2909824.3020222

2. RELATED WORK

Numerous studies have investigated the effects of social robots in tutoring scenarios. Empirical evidence demonstrates that learning with and from a physically present, interactive robot can be more effective than learning from classical on-screen media [14, 15, 20], and that robots can help children to develop in educational settings [10, 18, 24, 27]. However, at the same time, it is found that suboptimal behaviour of the robot can hamper learning [17]. Thus, a crucial ingredient for successful robot tutoring is the ability to provide personalized lessons [22] and to adapt in appropriate ways to the needs of the learner. The key question is when and how to adapt robot tutoring, according to which adaptation strategies, and based on what features of the state of learner or the tutoring interaction.

2.1 Approaches to Adaptive Tutoring

In the realm of Intelligent Tutoring System (ITS), dedicated *pedagogical modules* are employed for planning an optimal path through the curriculum by using an internal model of the learner's present knowledge state (cf. [8]). Cakmak and Lopes [3], for example, proposed a teaching algorithm that selects the most informative demonstrations for the learner. This learning agent makes use of Inverse Reinforcement Learning (IRL) to reduce the learner's hypothesis space of possible reward functions as fast as possible. In an evaluation, the authors showed that a learner trained with non-optimal selected expert demonstrations require significantly more demonstrations to achieve a similar performance as the optimally taught learner. This system, however, is designed for a sequential decision task in which no uncertainty about the learner's knowledge/skill exists. This assumption does not hold for the domain of L2 learning where the learner's current state of knowledge can, at best, be inferred from observed behaviour. Another important limitation of this approach is a lack of flexibility as no adaptation towards students' individual needs is considered.

Addressing especially the issue of adaptation towards students' individual needs, Partial Observable Markov Decision Processes (POMDPs) have been employed as basis for the pedagogical module of an ITS. Rafferty et al. [25], for instance, proposed different algorithms for planning an action-policy based on a POMDP and compared these against two different random and a maximum information gain (MIG) choice. They showed that even a simple action-policy based on a POMDP can achieve a significant faster skill learning than choosing actions randomly. But compared to the simple MIG algorithm, no significant difference was observed. Only with increasing skill space the MIG algorithm seems not to be sufficient anymore. A likely explanation for this finding is that the knowledge tracing model is insufficient. In addition, finding a good policy based on a POMDP is often computational intractable.

Clement et al. [4] compared two algorithms choosing the next skill and action in a tutoring interaction against a lesson given by a human expert. Both algorithms based on prior knowledge, e.g. the impact of actions on the learning gain or the difficulty of different types of tasks, which had been annotated by experts beforehand. The algorithms differed with regard to the adaptation method and the amount of additional knowledge stored besides the prior. The authors showed that even if the ITS does not make use of an internal model to store beliefs about the child's knowledge state regarding a specific skill, the use of their algorithm can lead to a higher learning gain compared to an expert lesson. Furthermore their second proposed algorithm, which additionally stores information about the knowledge state of the child, performed even better. Clement et al. concluded that extending their system with a more complex model for tracing the knowledge state of a student might lead to a higher learning.

An often criticized issue in this line of research is the lack of an effective *knowledge-tracing* method in the pedagogical module of an ITS that could be profitable for the learning interaction, e.g. by increasing the students' learning gain. Hence, we review research on knowledge tracing methods in the following.

2.2 Knowledge Tracing

Knowledge tracing aims to model learners' mastery of the knowledge being tutored. An often used approach is *Bayesian Knowledge Tracing* (BKT). BKT is a specific type of *Dynamic Bayesian Networks* (DBN), or more precisely, of Hidden Markov Models consisting of observed and latent variables. The latent variables represent the 'skills' and are classically assumed to be binary. That is, a skill is represented to be mastered or not. Generally, separate BKT networks are used for each skill to be learned [5]. Belief update is based on the observation of an answer to a given task testing a specific skill. The observed answer is binary too. Further, BKT models have two types of parameters: The emission probability and the transition probability. The emission probabilities are given by the 'guess probability' $p(guess)$, the probability of answering correctly without knowing the skill, and the 'slip probability' $p(slip)$ of answering wrongly although knowing the skill. In contrast, the transition probabilities are given by $p(t)$, the skill transition from unknown to known, and $p(f)$ the probability of forgetting a previously known skill. Often $p(f)$ is assumed to be zero.

Spaulding et al. [29] recently adopted BKT to trace the language-reading skill of children in robot-based language tutoring. They proposed the 'Affective BKT model', which is characterized by two further observable variables called 'smile' and 'engagement' to take into account the affective state of the child. This model structure allows emotions to influence the belief-state of each skill as they are included in every belief-update. The authors showed that the affective state of the children can be successfully integrated into BKT and that this approach outperforms traditional models for tracing the knowledge state in learning situations [29].

Another modification of BKT was published by Käser et al. [16]. Instead of using a dedicated BKT for every skill, they defined one comprehensive DBN to trace the knowledge on all skills to be learned. This enables to trace the knowledge on each skill individually and, in addition, to represent and reason with skill inter-dependencies. This allows for searching some kind of order in which skills may be learned best. The authors could demonstrate that this more detailed model outperforms other traditional models of knowledge tracing, including the normal BKT, with regard to the accuracy of the skill belief [16].

Finally, Gordon et al. [11] recently presented a so-called 'active learner model' to trace the word-reading skills of small children. This model does not work on the basis of BKT but employs a simple distance metric to approximate the conditional probability $p(w_2|w_1)$ of whether the child

can read a word w_2 if it already knows the word w_1. Their evaluation showed that their system is able to adapt to users of different age and to trace their reading knowledge up to a certain extent [11].

In this paper we present an expandable model based on BKT for knowledge tracing that, in contrast to the systems reviewed above [16, 29], allows for the simulation of actions and decision-making in teaching interactions.

3. ADAPTIVE LANGUAGE TUTORING

As a basis for our approach to adaptive language tutoring, we adopt the Bayesian Knowledge Tracing model [5] which has been successfully employed in other work and was shown to be easily extensible. However, we modify and extend the BKT model in order to enable predictive decision-making based on the represented beliefs about the learner's knowledge state. In this section, we first introduce our version of BKT and then present the approach for decision-making.

3.1 Bayesian Knowledge Tracing

The traditional approach to BKT uses only one latent variable S to represent the skill belief and one observable variable O for the user's answer. This suffices to represent if a skill is mastered or not, and how probable it would be that the user will answer correctly. Also, this information can be used to choose the next skill to learn, e.g. the skill which has the lowest belief probability of having been mastered. However, this model does not include information about how a skill can be addressed for teaching. In consequence, there is no possibility to take possible actions and their influence on the update of skill beliefs into account. We thus add a decision node A for actions to the Bayesian network (see Figure 1). This node not only influences the possible observation but also the belief update in the next time step. Further, we use a latent variable S that can attain six discrete values for each skill, corresponding to six bins for the belief state (0%, 20%, 40%, 60%, 80%, 100%). This allows for a more detailed model of the impact of tutoring actions on the possible observations and skills. Moreover, it becomes possible to better quantify the robot's uncertainty about the learner's skill.

With these changes, especially the conditional probability table $p(O^t|S^t)$ and the additional influence of the action A^t on the observable (now $p(O^t|S^t, A^t)$), the classical BKT update function, which was based on simple assumptions about guessing $p(guess)$ and slipping $p(slip)$ during the answer process, cannot be applied anymore. Instead, we apply a normal Bayesian update rule for the conditioning of skill beliefs including a transition probability $p(S_i^{t+1}|s_k, O^t, A^t)$ where s_k identifies a bin of the skill S_i^t. As a simplification we substitute this probability with $p(S^{t+1}|s_k)$:

$$\begin{aligned} p(S_i^{t+1}) &:= p(S_i^{t+1}|O^t, A^t) \\ &= \sum_{s_k \in S_i^t} [p(S_i^{t+1}|s_k, O^t, A^t) \cdot p(s_k|O^t, A^t)] \\ &\approx \sum_{s_k \in S_i^t} [\frac{p(O^t|s_k, A^t) \cdot p(A^t|s_k) \cdot p(s_k)}{p(O^t, A^t)} \cdot p(S_i^{t+1}|s_k)] \end{aligned}$$

3.2 Predictive Decision-Making

The extended BKT model is used to decide which tutoring action the robot should take next. At first, the skill to address with the next tutoring action is chosen. For this, the Kullback-Leibler divergence (KLD) between the current skill belief and the desired skill belief is used, the latter being a maximally certain belief in a maximally high skill of the learner:

$$next\_skill = \underset{\forall S_i^t \in \mathbb{S}}{\text{argmin}}[\alpha(S_i^t) \cdot KLD(p(S_i^t), p(S_{opt}))]$$

$\mathbb{S}$ represents the set of all skills that can be addressed, which consists of all words to be taught to the user. $p(S_{opt})$ is the desired belief for each skill, which means 99.999% of probability mass in the last bin (100%). The factor $\alpha(S_i^t)$ has been added for each skill to regulate the skill occurrence frequency. It is decreased each time the skill is addressed, and it is increased if another skill is being practised. In this way, the skill-selection algorithm takes care of the maximization of each skill belief as well as the balancing of all skills.

After the skill has been chosen, the next step is to decide with which tutoring action this should be done. Here, we consider abstract tasks as tutoring actions. These tasks will have to be mapped onto concrete exercises or pedagogical acts at a later stage in the robot control architecture (see Section 4). For simplicity, we distinguish between tutoring actions according to the difficulty (easy, medium or hard) of the task that addresses the corresponding skill. Finding the best action a_l for a given skill S_i^t is thus a minimization problem of the following form:

$$next\_action = \underset{\forall a_l \in A^t}{\text{argmin}}[\alpha(a_l) \cdot KLD(p(S_i^{t+1}), p(S_{opt}))]$$

where

$$\begin{aligned} p(S_i^{t+1}) &:= p(S_i^{t+1}|a_l) \\ &= \sum_{s_k \in S_i^t} \sum_{o_j \in O^t} p(S_i^{t+1}|o_j, s_k, a_l) \cdot p(o_j, s_k|a_l) \\ &\approx \sum_{s_k \in S_i^t} p(s_k|a_l) \sum_{o_j \in O^t} p(S_i^{t+1}|s_k) \cdot p(o_j|s_k, a_l) \end{aligned}$$

with

$$p(o_j, s_k|a_l) = p(o_j|s_k, a_l) \cdot p(s_k|a_l)$$

Here, $p(S_i^{t+1})$ could be seen as predicting the effect of applying the current action a_l to the skill S_i, where we again

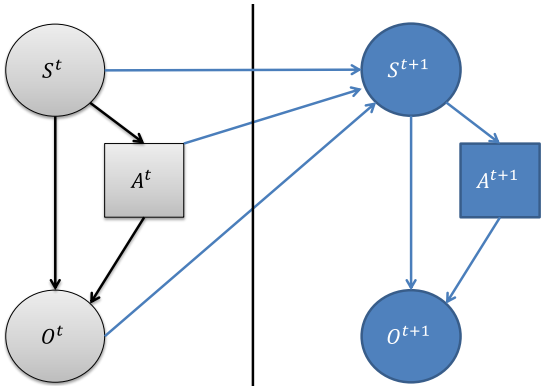


Figure 1: Dynamic Bayesian Network for BKT: The action node A^t predicts the observation O^t and influences the belief update of S^t for the next time step $t+1$.

substitute the transition probability $p(S_i^{t+1}|o_j, s_k, a_l)$ with $p(S_i^{t+1}|s_k)$ regarding simplicity. In addition, here again the skill belief is compared with $p(S_{opt})$ which represents the desired tutor belief state for each skill. The factor $\alpha(a_l)$ provides a more detailed selection of the "best" action. This way, the model will select an easy task if the skill is believed to be low, a hard task if it is high, and medium in-between. The goal of this strategy is to create a feeling of flow which can lead to better learning results [2, 7, 12]. Thus, it strives not to overburden the learner with too difficult tasks nor to bore him with too easy tasks, both of which may lead to frustration and thus hamper the learning [9, 13].

4. ROBOT LANGUAGE TUTORING

The adaptive model as described in the previous section has been brought to application in a child-robot second language (L2) tutoring game on the basis of empirical data from adult-child language tutoring interactions.

4.1 Empirical Basis

To design a tutoring interaction that matches children's needs, we decided to design the interaction on an empirical basis of language tutoring data. We collected video recordings of language tutoring games as they take place in kindergartens. Given that 1:1 interactions of educator and child can hardly be realized in kindergartens, the games typically involve one educator and a small group of children. Data of four language tutoring games have been collected: reading a picture book together with children in an interactive manner; card game "I spy with my little eye"; card game "I'm giving you a present"; and a rhyming game. The children were between four and six years of age. The data collected comprises round about 681 min of video data. These video data have been transcribed and annotated with regard to the following categories:

- **Dialogue acts**: Utterances are classified due to the underlying intention based on the DAMSL annotation scheme [6].
- **Children's mistakes**: Types of language errors the children made, e.g. wrong plural form, missing articles, wrong syntax, etc.
- **Educator's speech repair**: Pedagogical acts used to correct the errors, e.g. reformulation, corrected repetition, etc.
- **Nonverbal behaviour**: Nods, smiles, gestures etc. used by the educators.

On the basis of these annotations, we identified some overall patterns to inform the detailed design of the robot's behaviour. These fall basically into two categories, (i) overall interaction structure and (ii) feedback behaviour by the educators.

4.1.1 Overall Interaction Structure

A common pattern in all language tutoring games under investigation was the following basic course of actions:

1. **Opening:** Marks the beginning of the interaction and should mitigate the children's timidity as well as it should motivate the child.
2. **Game Setup:** This step is used to prepare the game by explaining the task and clarify the necessary terms.
3. **Test run:** A test run of the game is conducted to test whether the instructions have been understood and to practice the game flow.
4. **Game:** Here, the main interaction game takes place. Every move is accompanied by the educator's feedback and motivations to continue.
5. **Closing:** Marks the end of the learning interaction. Additionally, it is used to ensure motivation for future interactions by acknowledge the participation, joint singing a goodbye song and an outlook on what's going to happen next time.

4.1.2 Educator's Feedback Behaviour

In addition, we analysed the educators' behaviour when providing children with feedback. An important and common pattern is that language errors are almost never corrected explicitly. Instead, feedback is always provided in a positive way, falling into one of the following categories with the percentage of their occurrence given in squared brackets: (i) **praising the child** for a correct utterance whereby praise is often combined with a repetition of the correct word [13%] (ii) **implicit correction** in case of an error made by the child: repetition of the word as if correct (e.g. correct pronunciation, with article, plural form, etc.) [54%], (iii) **correct recasting of a sentence**, e.g. after syntax errors [32%], (iv) **moving on to next task**, e.g. when the child's message is unclear due to incomprehensible pronunciation [1%]. All kinds of educators' feedback behaviour is typically accompanied by looking at the child, smiling and nodding.

4.2 Game Setup

We have chosen the game "I spy with my little eye..." as a paradigm for our child-robot language tutoring game. The robot – in the role of a tutor, assisting the child in learning novel L2 vocabularies – is acting as 'the spy'. The child-robot setting is further enriched with a tablet PC on which objects are displayed (see Figure 2). In addition, the tablet's touch-screen displays three buttons to enable further user input in terms of 'yes' and 'no' answers as well as the option to let the robot repeat its previous statement.

A basic move of the game is structured as follows: It starts with a set of objects being displayed on the tablet screen and the robot saying "I spy with my little eye, something that is ...", followed by a foreign language word that refers to a property of one of the items on the screen. The child's task is now to respond by selecting the object referred to via touch input on the tablet. The robot's feedback behaviour in response to a correct or false answer is realized on the basis of our empirical data (see Section 4.1.2). That is, the robot responds to correct answers by praising the learner as well as repeating the L2 word and the corresponding L1 translation. In case of a false guess by the child, the robot explains the correct meaning of the to-be-learned word one more time. In addition, the wrongly chosen object as well as the actually correct object are both displayed on the tablet screen and the child is asked to select the correct object. The overall game structure is framed by the other elements making up typical language tutoring games in adult-child interaction (see Section 4.1.1).

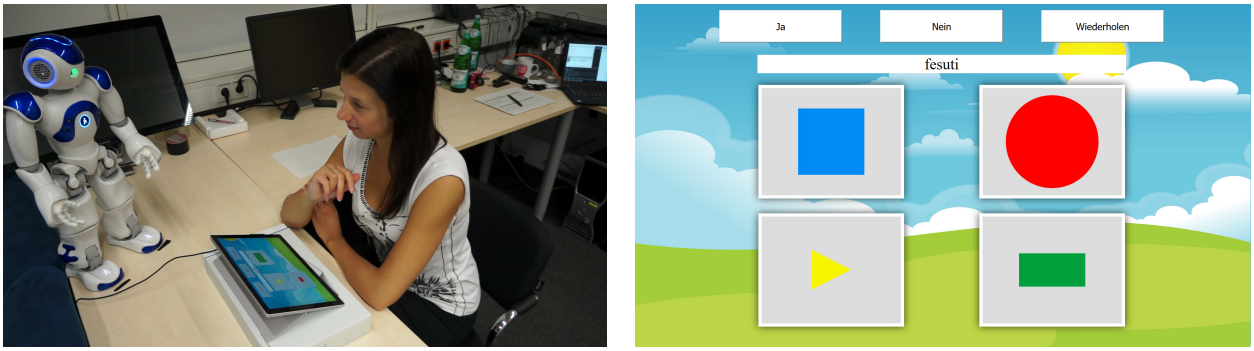


Figure 2: Experimental setup (left) with a participant sitting in front of a tablet displaying the graphical user interface (right). The robot *Nao* stands next to the tablet slightly rotated towards the user.

4.3 Technical Realization

We employed the Nao robot[2] for our language tutoring game. It is standing in a bit more than 90 degrees rotated, to the right of the participant. In addition a Microsoft Surface Pro 4[3] tablet PC is used to catch the user input and to display the graphical user interface realized via a HTML website. For the implementation of the interaction and dialogue structure, the state-chart based dialogue-manager IrisTK has been used [28]. NAOqi[4] has been applied as middleware between the robot, the graphical user interface, the dialogue manager, and our developed adaptive tutoring model. NAOqi is shipped with each Nao robot and allows to communicate via a simple event system between various programming-languages (Python, Java, C++, JScript).

5. EVALUATION STUDY

To assess the effects of our adaptive model on L2 word learning, we set up an evaluation study based on the language tutoring game described in the previous section. The major objective behind this study was to evaluate the effects of the adaptive model on learners' performance. We used the Nao robot to deliver all task information and direct feedback to the learner. This enables us to test the model within the desired final setting, including the effects of a robot's presence in the tutoring interaction. Given that children show a high degree of inter-individual variation and might further need child-specific adaptations of, for instance, synthesized speech to enable them to understand what the robot says, we decided to conduct this first study with adult learners.

We employed a between-subjects design with a manipulation of training type: Participants learned L2 vocabulary items either with the fully adaptive model, or in a random control condition. In the adaptive condition, the skill to be taught and the action to address the skill were chosen by the model as described in Section 3. In our language tutoring game, skill relates to the foreign language words and action refers to the specific task used in the game (target word, objects displayed). The difficulty of the actions/tasks in this study were implemented by using less or more distractor objects that were shown together with the correct object on the screen. For instance, an easy task consisted of two distractor objects, whereas a hard task had four distractors. Distractors were chosen with respect to the skill beliefs of the user, with the set of objects mainly consisting of items for which the L2 words were still/mostly unknown by the learner.

As shown by Craig et al. [7], better learning performance is to be expected when learners have to expend the right amount of cognitive effort (i.e. not too hard or too easy tasks). Accordingly, while learning with our model in the adaptive condition, no hard tasks are shown until the system believes the user to have basic knowledge on all skills. Then, the system will increase task difficulty (as determined by the adaptive tutoring model) by adding distractor objects. Note, however, that at a certain point the user will know too many skills/words so that finding a distractor set (i.e. task difficulty) that cannot be sorted out by exclusion becomes impossible. In the control condition, all skills are taught in a random order and always with 'medium' task difficulty.

Participants' performance was assessed with two measures: (1) we tracked learners' response behaviour over the course of the training to investigate the progress of learning, (2) we conducted a post-test on the taught vocabulary in the form of both L1-to-L2 translations and L2-to-L1 translations to assess participants' state of knowledge subsequent to the intervention.

5.1 Materials

The training materials for the study comprised German–'Vimmi' word pairs. Vimmi is an artificial language created for experimental purposes [23] that aims to avoid associations with other known words or languages. The Vimmi items are created according to Italian phonotactic rules. Ten items have been chosen: four colour terms, four shape-encoding terms and two terms describing size (see Table 1).

5.2 Procedure

Upon entering the lab, participants were randomly assigned to one of the two experimental conditions. They were informed that they take part in an experiment on foreign language learning and were asked to sign an informed consent form. They also filled out a questionnaire that covered personal information like age and nationality as well as a personal estimation of language learning skills in general and memorization ability for L2 vocabulary.

[2]https://www.ald.softbankrobotics.com/en/cool-robots/nao
[3]https://www.microsoft.com/surface/en-gb/devices/surface-pro-4
[4]http://doc.aldebaran.com/2-1/naoqi/

N	German	Vimmi	English translation
1	blau	bati	blue
2	grün	uteli	green
3	gelb	dirube	yellow
4	rot	fesuti	red
5	rund	beropuga	round
6	dreieckig	pewo	triangular
7	quadratisch	tanedila	square
8	rechteckig	paltra	rectangular
9	klein	kiale	small
10	groß	ilado	big

Table 1: The 10 words from Vimmi to be learned in the evaluation study with its corresponding translation in German as well as English for comprehension purposes.

Next, a list of the to-be-learned Vimmi items were presented to the participants for 30 seconds. This was to enable participants to practice the items right from the first game interaction on. Then, the learning interaction with the Nao robot began. After introducing itself, the robot explained the "I spy with my little eye"-game and started a test-run with the participants. Once this test run was finished and the participants agreed that (s)he understood the game, the main interaction consisting of a total of 30 trials (game moves) began. Each trial addressed one vocabulary item as described in Section 4.2. That is, the robot asked for one of the objects displayed on the tablet screen, whereby the question was in L1 (German) for the most part, except for the referring, to-be-learned word in L2 (Vimmi). After 30 trials, the game was finished, the Nao robot thanked the participants and said goodbye.

Subsequent to the interaction with the robot, participants' learning performance was assessed with a post-test. In an interview with the experimenter, they had to translate the ten to-be-learned vocabulary items from German to Vimmi and likewise from Vimmi to German (both in randomized order). The whole interaction and the vocabulary-post-test at the end of the study were recorded with an external camera. Also the system decisions taken during the interaction and the probability distributions for each updated skill belief were logged.

5.3 Participants

A total of 40 participants (20 per condition) with an average age of 24.13 ($SD = 3.82$) took part in this study (16 males and 24 females). All participants had very good command of the German language and normal or corrected sight. All of them were paid or received credits for their participation.

5.4 Results

5.4.1 *Learning Progress During Training*

In order to assess the learners' progress during training, we compared the number of correct responses addressing the initial quarter of the tutoring game (first seven items) against the final quarter (last seven items). When an item occurred repeatedly within the initial quarter, the first occurrence has been taken into account. When an item occurred repeatedly within the final quarter, the last occurrence has been considered.

	Adaptive (A)		Control (C)		A, C	
	M	SD	M	SD	M	SD
F7	3.75	1.37	4.00	1.17	3.88	1.27
L7	6.90	0.31	5.15	1.69	6.03	1.49
F7, L7	5.33	0.69	4.58	1.12		

Table 2: Means (M) and standard deviations (SD) of correct answers for the initial quarter of the training interaction (first seven items – F7) and the final quarter (last seven items – L7) in each condition, as well as the inter-model (A, C) and intra-model (F7, L7) means and standard deviations.

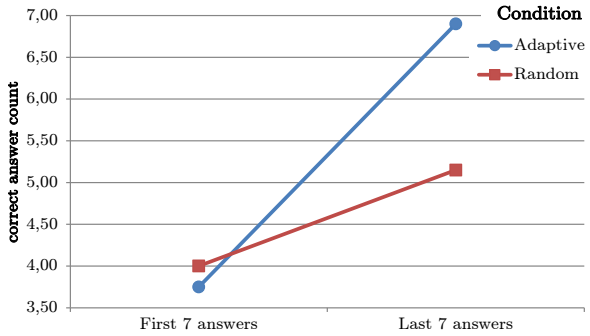


Figure 3: Mean numbers of correct answers at the beginning (first 7) and end (last 7) of the interaction in the different conditions.

A mixed-design ANOVA with training phase (initial, final) as a within-subjects factor and training type (adaptive-model-based, control) as between-subjects factor has been conducted. Results are summarized in Table 2 and Figure 3. Not surprisingly, there was a main effect of training phase at a significant level ($F(1, 38) = 66.85, p < .001, \eta^2 = .64$): Learners' performance was significantly better in the final phase as compared to the initial phase. In the first quarter of training, participants achieved a mean of 3.88 ($SD = 1.27$) correct responses, whereas in the final quarter, a mean of 6.03 ($SD = 1.49$) items was selected correctly. More interestingly, there was also a main effect of training type ($F(1, 38) = 6.52, p = .02, \eta^2 = .15$) such that participants who learned in the adaptive condition had a higher score of correct answers ($M = 5.33, SD = .69$) as compared to learners in the control condition with an average of $M = 4.58$ ($SD = 1.12$) correct answers. Finally, the interaction between training phase and training type was also significant ($F(1, 38) = 14.46, p = .001, \eta^2 = .28$) indicating that the benefit of adaptive-model-based training develops over time (see Figure 3). While participants' response behaviour in the first quarter of training was similar across conditions, a benefit of training with the adaptive model became evident in the final quarter. At this stage of training, participants in the adaptive model condition achieved a mean of $M = 6.9$ ($SD = .31$) correct responses, whereas participants in the control condition achieved a mean of $M = 5.15$ ($SD = 1.69$) correct responses.

	Adaptive (A)		Control (C)	
	M	SD	M	SD
German-to-Vimmi	3.95	2.56	3.35	1.98
Vimmi-to-German	7.05	2.56	6.85	2.48

Table 3: Results of both post-tests (German-to-Vimmi and Vimmi-to-German): Means (M) and standard deviation (SD) of correct answers grouped by the experimental conditions.

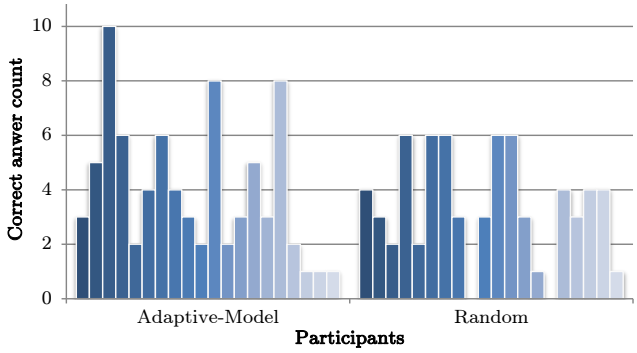


Figure 4: Participant-wise amount of correct answers grouped by the different conditions for the German-to-Vimmi post-test.

5.4.2 *Post-Test*

Participants' learning performance subsequent to the intervention has been measured with two translation tests (L2-to-L1 and L1-to-L2). Results are summarized in Table 3. Paired-samples t-tests were conducted to compare the number of correctly recalled words after training with the adaptive model as compared to training in the control condition. For the German-to-Vimmi translation, there was no significant main effect ($T(38) = .25, p = .80$). Participants who trained with the adaptive-model recalled a mean of 3.95 ($SD = 2.56$) out of ten words correctly, while participants in the control condition recalled a mean of 3.35 ($SD = 1.98$) words. Likewise, there was no significant main effect ($T(38) = .83, p = .41$) for the Vimmi-to-German translation task. Participants' performance after learning with the adaptive model amounted to a mean of 7.05 ($SD = 2.56$) correct items, participants' performance in the control condition to a mean of 6.85 ($SD = 2.48$) correct items.

Although no main effect of training type emerged in the post-test, some details might nevertheless be worth mentioning. In the German-to-Vimmi post-test, a maximum of ten correct responses was achieved by participants in the adaptive-model condition, whereas the maximum on the control condition were six correct answers. Moreover, there were two participants in the control condition who did not manage to perform any German-to-Vimmi translation correctly. In the adaptive-model condition, all participants achieved at least one correct response (see Figure 4).

6. CONCLUSION

In this paper we have presented a novel approach to personalize language tutoring in human-robot interaction. This adaptive tutoring is enabled through a model of how tutors mentalize about learners – by keeping track of their knowledge state and by selecting the next tutoring actions based on their likely effects on the learner. This is realized via an extended model that combines Bayesian Knowledge Tracing (of the learned) with tutoring actions (of the tutor) in one causal probabilistic model. This allows, for selecting skills and actions based on notions of optimality – here the desired learner's knowledge state as well as optimal task difficulty – to achieve this for a given skill. This model has been implemented into a robot language tutoring game and tested in a first evaluation study.

The analysis of participants' response behaviour over the course of training has clearly shown that participants learned the L2 words during the human-robot interaction. Importantly, they learned more successfully with our adaptive model as compared to a randomized training. That is, the repeated trials addressing still unknown items as chosen by the adaptive model (until the belief state about these words equalled that of known items) outperformed the tutoring of the same material (same number of trials and items) but in randomized order. In the post-test, however, there was no significant difference across experimental conditions, despite a trend towards better performance in the adaptive model conditions over the controls.

Different explanations may account for this inconsistent finding. One potential explanation could be that the way how responses were prompted was not identical in training sessions and post-test. In the training sessions, participants saw pictures reflecting the meaning of the to-be-learned words whereas in the post-test they just received a linguistic cue in form of a word they had to translate. It might be that repeated trials as they were particularly supported for difficult-to-remember items by the adaptive model, led to stronger associations between linguistic and imagistic materials. This might have caused a stronger decline of correct responses for participants who trained with the adaptive model as opposed to those in the control condition. An alternative explanation could be that test results measured immediately after the training session are subject to strong inter-individual differences among learners. This is the reason why studies on vocabulary learning usually range over repeated sessions spread over several days (cf. 1). A typical pattern is that significant results emerge after two or three sessions/days and/or in the long-term (measured several weeks after training took place). So it might well be that further training sessions or delayed tests might result in a post-test performance that matches the picture of the during-session performance.

One might argue that the performance of our adaptive model is comparable to the vocabulary learning technique of *spaced repetition* as implemented, for instance, in the Leitner system [21]. In this system flashcards are sorted into groups according to how well the learner knows each one. Learners try to recall items written on a flashcard. If they succeed, the card is sent to the next group. If they fail, the card is sent back to the first group. Each succeeding group has a longer period of time before the learner is required to revisit the cards. This way all items, that are hard to remember for the learner will be repeated more often. In contrast to such spaced repetition systems, our model is more flexible as it can vary the difficulty of the tasks by providing more or less distractor items. In addition, we plan a more comprehensive action space of the model to account for motivating actions

where necessary or adaptations in the robot's verbal or non-verbal behaviour.

Overall, results from the evaluation study are, at least, in parts very promising: learners' performance during training was significantly improved by personalized robot tutoring based on the adaptive model. Nevertheless, the fact that this positive effect did not hold in the post-test, inter alia, marks a starting point for further refinements of the model: Training stimuli should be designed such that they match the way language learners need to apply them best possible. That is, when the aim is to enable people to translate words from one language to another, training stimuli should provide cues for this process of mapping linguistic materials on each other. Moreover, a further study with more learning sessions (e.g. over several days as common in many vocabulary studies) should be conducted. Regarding the model itself, we plan to incorporate skill-interdependencies as well as to take the affective user state into account, too. Both kind of extensions have been shown to improve learning [16, 29]. Additionally, the model can (and is meant to) provide a basis for exploiting the full potential of an embodied tutoring agent. Regarding this, we plan to advance the model such that the robot's verbal and non-verbal communicative behaviour is adapted to the learner's state of knowledge and progress. Specifically, we aim to enable dynamic adaption of (i) embodied behaviour such as iconic gesture use to be known to support vocabulary acquisition as a function of individual differences across children (cf. [26]); (ii) the robot's synthetic voice to enhance comprehensibility and prosodic focusing of content when needed; and (iii) the robot's socio-emotional behaviour depending on the learners' current level of motivation or engagement. Further, as the long-term goal of our work is to enable robot-supported language learning for pre-school children, another important goal is to make children-specific adaptations to the language game and test it in child-robot interaction studies.

7. ACKNOWLEDGEMENTS

This work was supported by the L2TOR (www.l2tor.eu) project supported by the EU Horizon 2020 Program, grant number: 688014, and by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

8. REFERENCES

[1] K. Bergmann and M. Macedonia. *A Virtual Agent as Vocabulary Trainer: Iconic Gestures Help to Improve Learners' Memory Performance*, pages 139–148. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[2] D. E. Berlyne. Conflict, arousal, and curiosity. 1960.

[3] M. Cakmak and M. Lopes. Algorithmic and human teaching of sequential decision tasks. In *AAAI Conference on Artificial Intelligence (AAAI-12)*, 2012.

[4] B. Clement, D. Roy, P.-Y. Oudeyer, and M. Lopes. Multi-armed bandits for intelligent tutoring systems. *arXiv preprint arXiv:1310.3174*, 2013.

[5] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[6] M. G. Core and J. Allen. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56. Boston, MA, 1997.

[7] S. Craig, A. Graesser, J. Sullins, and B. Gholson. Affect and learning: An exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, 29(3):241–250, 2004.

[8] C. Dede. A review and synthesis of recent research in intelligent computer-assisted instruction. *International Journal of Man-Machine Studies*, 24(4):329–353, 1986.

[9] S. Engeser and F. Rheinberg. Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion*, 32(3):158–172, 2008.

[10] M. Fridin. Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education. *Comput. Educ.*, 70:53–64, Jan 2014.

[11] G. Gordon and C. Breazeal. Bayesian active learning-based robot tutor for children's word-reading skills. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 1343–1349. AAAI Press, 2015.

[12] J. Gottlieb, P.-Y. Oudeyer, M. Lopes, and A. Baranes. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593, 2013.

[13] M. J. Habgood and S. E. Ainsworth. Motivating children to learn effectively: Exploring the value of intrinsic integration in educational games. *The Journal of the Learning Sciences*, 20(2):169–206, 2011.

[14] J. Han, M. Jo, S. Park, and S. Kim. The educational use of home robots for children. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 378–383, Aug 2005.

[15] E. ja Hyun, S. yeon Kim, S. Jang, and S. Park. Comparative study of effects of language instruction program using intelligence robot and multimedia on linguistic ability of young children. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, pages 187–192, Aug 2008.

[16] T. Käser, S. Klingler, A. G. Schwing, and M. Gross. *Beyond Knowledge Tracing: Modeling Skill Topologies with Bayesian Networks*, pages 188–198. Springer International Publishing, Cham, 2014.

[17] J. Kennedy, P. Baxter, and T. Belpaeme. The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 67–74, New York, NY, USA, 2015. ACM.

[18] J. Kory and C. Breazeal. Storytelling with robots: Learning companions for preschool children's language development. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 643–648, Aug 2014.

[19] J. Kory Westlund, G. Gordon, S. Spaulding, J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal. Learning a second language with a socially assistive robot. In *The 1st International Conference on Social Robots in Therapy and Education*, 2015.

[20] H. Kose-Bagci, E. Ferrari, K. Dautenhahn, D. S. Syrdal, and C. L. Nehaniv. Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*, 23(14):1951–1996, 2009.

[21] S. Leitner. So lernt man lernen: Der weg zum erfolg [learning to learn: The road to success]. *Freiburg: Herder*, 1972.

[22] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati. The physical presence of a robot tutor increases cognitive learning gains. In *CogSci*. Citeseer, 2012.

[23] M. Macedonia, K. Müller, and A. D. Friederici. Neural correlates of high performance in foreign language vocabulary learning. *Mind, Brain, and Education*, 4(3):125–134, 2010.

[24] J. R. Movellan, M. Eckhardt, M. Virnes, and A. Rodriguez. Sociable robot improves toddler vocabulary skills. In *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on*, pages 307–308, March 2009.

[25] A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching via pomdp planning. *Cognitive Science*, 2015.

[26] M. L. Rowe, R. D. Silverman, and B. E. Mullan. The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemporary Educational Psychology*, 38(2):109–117, 2013.

[27] M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse. 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*, CHI '10, pages 1613–1622, New York, NY, USA, 2010. ACM.

[28] G. Skantze and S. Al Moubayed. Iristk: A statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 69–76, New York, NY, USA, 2012. ACM.

[29] S. Spaulding, G. Gordon, and C. Breazeal. Affect-aware student models for robot tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, pages 864–872, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems.

Workshop on Robots for Learning - R4L

Wafa Johal
CHILI/LSRO Labs
École Polytechnique
Fédérale Lausanne
Lausanne, Switzerland
wafa.johal@epfl.ch

Paul Vogt
Tilburg center for Cognition
and Computation
Tilburg University
Netherlands
p.a.vogt@uvt.nl

James Kennedy
Centre for Robotics and
Neural Systems
Plymouth University
United Kingdom
james.kennedy
@plymouth.ac.uk

Mirjam de Haas
Tilburg center for Cognition
and Computation
Tilburg University
Netherlands
mirjam.dehaas@uvt.nl

Ana Paiva
Instituto Superior Técnico
University of Lisbon
Portugal
ana.paiva@inesc-id.pt

Ginevra Castellano
Department of Information
Technology
Uppsala University
Sweden
ginevra.castellano
@it.uu.se

ABSTRACT

While robots have been popular as a tool for STEM teaching, the use of robots in other learning scenarios is novel. The field of HRI has started to report on how to make effective robots usable in educational contexts. However, many challenges remain. For instance, which interaction strategies aid learning, and which hamper learning? How can we deal with the current technical limitations of robots? Answering these and other questions requires a multidisciplinary effort, including contributions from pedagogy, developmental psychology, (computational) linguistics, artificial intelligence and HRI, among others. This abstract provides a brief overview of the current state-of-the-art in social robots designed for learning and describes the aims of the Robots for Learning (R4L) workshop in bringing together a multidisciplinary audience for furthering the development of market-ready educational robots.

Keywords

Human-Robot Interaction; Robots in Education; Tutor Robots; Child-Robot Interaction

1. INTRODUCTION & BACKGROUND

An increasing amount of Human-Robot Interaction (HRI) research is focused on the development of applications of service robots in everyday life. In education, while robots have been popular as a focus for STEM teaching (cf. Lego Mindstorms or Thymio [7]), the use of robots in other learning scenarios is novel.

Mubin et al. [5] distinguish three roles for robots in education: 1) **tutors** - providing help to students, 2) **peers** - stimulating learning, and 3) **tools** - physically enhancing a concept to learn. In the 1970's and 80's robots tended to be introduced in schools as a tool for teachers to teach robotics or other STEM subjects. However, this specificity of robot usage penalized their adoption in educational contexts [2]. Nowadays, with robots being cheaper and more easily deployable, application in education becomes possible for other types of learning.

The field of HRI has started reporting on how to make effective robots and how to measure their efficacy [3, 8]. Robots have the potential to enhance learning via kinesthetic interaction, can improve the learner's self-esteem, and can provide empathic feedback [1, 4, 9]. Finally, robots have been shown to engage the learner, to motivate her in the learning task or to stimulate collaboration in a group [6]. However, many challenges remain and this workshop aims to bring together a multidisciplinary group of researchers to discuss these challenges and share expertise. Such challenges and questions that are yet to be comprehensively addressed by the research community include: the effective involvement of education practitioners in the design of activities, the outcome of long-term learning with robots, appropriate educational strategies for use in HRI, and the influence of HRI on affective aspects of learning, such as motivation and self-efficacy.

The second iteration of this workshop builds on the previous version hosted at the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2016. The previous workshop utilised keynote speakers, participant speakers, and small group discussions to raise issues and challenges facing the community researching robots for use in delivering educational content. The second version of this workshop seeks to engage with more researchers in the field, and draw a more multidisciplinary audience to further the development of market-ready educational robots.

2. OUTLINE OF THE WORKSHOP

The aim of this workshop is to engage scholars who wish to gain expertise in education and in robotics. Participants will benefit from hearing from the forefront of field and

HRI '17 Companion March 06-09, 2017, Vienna, Austria

ACM ISBN 978-1-4503-4885-0/17/03.

DOI: http://dx.doi.org/10.1145/3029798.3029801

from discussions on how to move from fundamental research towards the development of market-ready educational robots.

The workshop aims will be achieved through presentations and discussions. Prospective participants are invited to submit 4-6 page papers describing work in progress, or containing preliminary results to discuss with the community. In order to stimulate interactions, the workshop will include short position paper presentations and poster sessions. The afternoon will be dedicated to discussion, including both a panel session and semi-structured group discussions.

3. ORGANIZERS

Wafa Johal, PhD. is a postdoctoral researcher within the CoWriter and Cellulo projects dealing with robots for education in the CHILI and LSRO Labs at EPFL. She obtained her PhD in 2015 from the University of Grenoble (France) focusing on body signals in Child-Robot Interaction.

Paul Vogt is Associate Professor in Language learning and HRI. He is a trained cognitive scientist and holds a PhD in Artificial Intelligence. His research focuses on 1st and 2nd language acquisition using methods from ethnographic research, psycholinguistics, computational modelling of language acquisition and HRI. Paul is one of the principal investigators in the L2TOR project.

James Kennedy is currently completing his PhD in Human-Robot Interaction at Plymouth University (U.K.). His research interests centre around social companion robots, particularly in educational interactions with children. He has been involved with the ALIZ-E, DREAM and L2TOR European projects.

Mirjam de Haas finished her Master's degree in Artificial Intelligence and is a PhD student in the L2TOR project. Her research focuses on the interaction between robots and children and how to design a child-friendly robot.

Ana Maria Paiva's main scientific interests lie in the area of Autonomous Agents, Embodied Conversational Agents and Robots, and Multiagent Simulation Systems. She has been researching in the area of artificial intelligence for the past twenty years. She is the principal investigator of the eCUTE project aiming to explore technologically-enhanced learning approaches for inter-cultural understanding.

Ginevra Castellano is an associate senior lecturer in intelligent interactive systems at Uppsala University, where she leads the Social Robotics Lab. She was the coordinator of the EMOTE project, which developed educational robots to support teachers in a classroom environment.

Sandra Okita is an Associate Professor of Technology and Education at Teachers College, Columbia University. Her current research interest is focused on the learning partnership between individuals and technology, and how technology intersects with learning and instructional processes.

Fumihide Tanaka, PhD, has been actively working in the area of educational robots and child-robot interaction, and is now recognized as one of the pioneers in this research area. He moved to academia in 2008, the University of Tokyo (2014), and is currently at the University of Tsukuba, Japan.

Tony Belpaeme's research focuses on cognitive robotics and social Human-Robot Interaction, in which natural and artificial cognition is considered to be closely intertwined with social interaction. He coordinates the L2TOR project on learning a 2nd language using robot tutors, and collaborates on several international research projects on HRI and cognitive robotics.

Pierre Dillenbourg is a former elementary school teacher. He graduated in educational science (University of Mons, Belgium). His research on learning technologies started in 1984. He obtained a PhD in computer science from the University of Lancaster (UK), in artificial intelligence applications for educational software. He is currently full professor in learning technologies, head of the CHILI Lab involved in both CoWriter and Cellulo projects.

4. ACKNOWLEDGMENTS

We would like to thank the Swiss National Science Foundation National Centre of Competence in Research Robotics and the EU H2020 L2TOR project (grant no. 688014).

5. ADDITIONAL AUTHORS

- Sandra Okita, Teachers College - Columbia University, United States, `okita@tc.columbia.edu`
- Fumihide Tanaka, University of Tsukuba, Japan, `tanaka@iit.tsukuba.ac.jp`
- Tony Belpaeme, Centre for Robotics and Neural Systems, Plymouth University, U.K. and Ghent University, Belgium, `tony.belpaeme@plymouth.ac.uk`
- Pierre Dillenbourg, CHILI Lab, École Fédérale Polytechnique Lausanne, Switzerland, `pierre.dillenbourg@epfl.ch`

6. REFERENCES

[1] G. Castellano, A. Paiva, A. Kappas, R. Aylett, H. Hastie, W. Barendregt, F. Nabais, and S. Bull. Towards empathic virtual and robotic tutors. In *Artificial Intelligence in Education*, pages 733–736. Springer, 2013.

[2] W. Gander, A. Petit, G. Berry, B. Demo, J. Vahrenhold, A. McGettrick, R. Boyle, M. Drechsler, A. Mendelson, C. Stephenson, C. Ghezzi, and B. Meyer. Informatics education: Europe cannot afford to miss the boat, Report of the joint Informatics Europe & ACM Europe Working Group on Informatics Education. Available at: http://europe.acm.org/iereport/ie.html, 2013.

[3] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme. Social Robot Tutoring for Child Second Language Learning. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, pages 67–74. ACM, 2016.

[4] S. Lemaignan et al. Learning by Teaching a Robot: The Case of Handwriting. *IEEE Robotics Automation Magazine*, 23(2):56–66, 2016.

[5] O. Mubin, C. J. Stevens, S. Shahid, A. A. Mahmud, and J.-J. Dong. A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, 1:209–0015, 2013.

[6] H. W. Park and A. Howard. Providing tablets as collaborative-task workspace for human-robot interaction. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 207–208. IEEE Press, 2013.

[7] F. Riedo, P. Rétornaz, L. Bergeron, N. Nyffeler, and F. Mondada. A Two Years Informal Learning Experience Using the Thymio Robot. In *Advances in Autonomous Mini Robots*, pages 37–48. Springer, 2012.

[8] F. Tanaka, K. Isshiki, F. Takahashi, M. Uekusa, R. Sei, and K. Hayashi. Pepper learns together with children: Development of an educational application. In *IEEE-RAS 15th International Conference on Humanoid Robots*, HUMANOIDS 2015, pages 270–275. IEEE, 2015.

[9] F. Tanaka and S. Matsuzoe. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, 1(1):78–95, 2012.

frontiers in Human Neuroscience

PERSPECTIVE
published: 02 March 2017
doi: 10.3389/fnhum.2017.00073

Child-Robot Interactions for Second Language Tutoring to Preschool Children

*Paul Vogt *, Mirjam de Haas, Chiara de Jong, Peta Baxter and Emiel Krahmer*

Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, Netherlands

In this digital age social robots will increasingly be used for educational purposes, such as second language tutoring. In this perspective article, we propose a number of design features to develop a child-friendly social robot that can effectively support children in second language learning, and we discuss some technical challenges for developing these. The features we propose include choices to develop the robot such that it can act as a peer to motivate the child during second language learning and build trust at the same time, while still being more knowledgeable than the child and scaffolding that knowledge in adult-like manner. We also believe that the first impressions children have about robots are crucial for them to build trust and common ground, which would support child-robot interactions in the long term. We therefore propose a strategy to introduce the robot in a safe way to toddlers. Other features relate to the ability to adapt to individual children's language proficiency, respond contingently, both temporally and semantically, establish joint attention, use meaningful gestures, provide effective feedback and monitor children's learning progress. Technical challenges we observe include automatic speech recognition (ASR) for children, reliable object recognition to facilitate semantic contingency and establishing joint attention, and developing human-like gestures with a robot that does not have the same morphology humans have. We briefly discuss an experiment in which we investigate how children respond to different forms of feedback the robot can give.

Keywords: social robots, second language tutoring, education, child-robot interaction, robot assisted language learning

OPEN ACCESS

Edited by:
Mila Vulchanova, Norwegian University of Science and Technology, Norway

Reviewed by:
Ramesh Kumar Mishra, University of Hyderabad, India
Vera Kempe, Abertay University, UK

****Correspondence:***
Paul Vogt
p.a.vogt@uvt.nl

Received: *26 October 2016*
Accepted: *06 February 2017*
Published: *02 March 2017*

Citation:
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. 11:73. doi: 10.3389/fnhum.2017.00073

SOCIAL ROBOTS FOR SECOND LANGUAGE TUTORING

Given the globalization of our society, it is becoming increasingly important for people to speak multiple languages. For instance, the ability to speak foreign languages fosters people's mobility and increases their chances for employment. Moreover, immigrants to a country need to learn the official host language. Since young children are most flexible at learning languages, starting second language (L2) learning in preschool would provide them a good opportunity to acquire the second language more fluently at a later age (Hoff, 2013).

One trend in the digital age of the 21st century is that technologies are being developed for educational purposes, including technologies to support L2 tutoring. There exist many forms of digital technologies for PCs, laptops or tablet computers that support second language learning, although there is little evidence about their efficacy (Golonka et al., 2014; Hsin et al., 2014). While children can benefit from playing with such technologies, these systems lack the situated and

embodied interactions that young children naturally engage in and learn from (Glenberg, 2010; Leyzberg et al., 2012). Social robots represent an emerging technology that provides situatedness and embodiment, and thus have potential benefits for educational purposes. In essence, social robots are autonomous physical agents, often with human-like feature, that can interact socially with humans in a semi-natural way for prolonged periods of time (Dautenhahn, 2007). The use of social robots, in comparison to more traditional digital technologies, allows for the development of tutoring systems more akin to human tutors, especially with respect to the situated and embodied social interactions between child and robot. Thus, this offers the opportunity to design robots such that they interact in a way that optimizes the child's language learning.

Recently, an increasing interest has emerged to develop social robots to support children with learning a second language (Kanda et al., 2004; Belpaeme et al., 2015; Kennedy et al., 2016). While a social robot cannot provide tutoring to the level humans can, recent studies suggest that using social robots can result in an increased learning gain compared to digital learning environments for tablets or computers (Han et al., 2008; Leyzberg et al., 2012). It is, however, unclear why this is the case. Perhaps the physical presence of the robot draws the attention of children for longer periods of time, but the embodiment and situatedness of the learning environment perhaps also helps the children to ground the language more strongly than interactions with virtual objects do.

While there is a fair body of research on robot tutors, a comprehensive description of the design features for a second language robot tutor based on what is known about children's language acquisition is lacking. What are the design features of child-robot interactions that would support second language learning? And, to what extent can these interactions be implemented in today's social robot technologies? In this perspective article, we try to answer these questions based on theoretical accounts from the literature on children's language acquisition in combination with our own experiences in designing a tutor robot.

DESIGNING CHILD-ROBOT INTERACTIONS

In our project, we aim to design a digital learning environment in which preschool children interact one-on-one with a social robot that supports either their learning of English as a foreign language, or the school language for those children who have a different native language (Belpaeme et al., 2015). In particular, the project aims to develop a series of tutoring sessions revolving around three increasingly complex domains (numbers, spatial relations and mental vocabulary). In each session, the child will engage with the robot (a Softbank Robotics NAO robot) in a game-like scenario focusing on learning a small number of target words. The contextual setting is generally displayed on a tablet computer that occasionally also provides some verbal support, however, the robot acts as the interactive tutor. Below we discuss the design features and considerations that we believe are crucial to design a successful tutoring system.

Peer-Like Tutoring

One of the first questions that comes up when designing a robot tutor is whether the robot should take the role of a teacher or a peer. Research on children's language acquisition has demonstrated that children learn more effectively from an adult who can use well-defined pedagogical methods for teaching children using clear directions, explanations and positive feedback methods (Matthews et al., 2007). However, designing and framing the robot as an adult tutor has the disadvantage that children will form expectations about the robot's behavior and proficiency that cannot be met with current technology (Kennedy et al., 2015). Due to technological limitations of the robot and underlying software, communication breakdowns are more likely to occur than with a human. For a peer robot introduced as a fellow language learner, breakdowns in communication are more acceptable. Moreover, interacting with robots acting as peers is conceived as more fun (Kanda et al., 2004), allows for learning-by-teaching (Tanaka and Matsuzoe, 2012) and has a proven to be efficient in teaching children how to write (Hood et al., 2015). Furthermore, there is some evidence that children's learning can benefit from interacting with peers (Mashburn et al., 2009). Given these considerations, we believe it is desirable to frame or introduce the robot as a peer and friend, yet design its interactions insofar possible based on pedagogically well-established strategies to scaffold language learning.

First Impressions

To implement effective tutoring, the robot needs to interact with children in multiple sessions, so they have to be motivated to engage in long-term interactions with the robot. Establishing common ground between child and robot can contribute to this (Kanda et al., 2004), but first impressions to establish trust and rapport are also crucial (Hancock et al., 2011).

Despite the wealth of studies regarding the introduction of entertainment robots as toys to children (e.g., Lund, 2003), surprisingly little research has been conducted on designing protocols on how to introduce a robot tutor to a group of preschool children. Fridin (2014) presents one exception, and found that introducing a robot tutor to children in group sessions improved subsequent interactions compared to introducing the robot to children in individual sessions. Another study by Westlund et al. (2016) found that the way a robot is framed, either as a machine or a social entity, affected the way children later engaged with the robot. They concluded that introducing the robot as a machine could create a more distant relation between child and robot, thus reducing acceptance. We therefore decided to frame the robot in our project as a social playmate for the children and introduced the robot in a group session. However, the NAO robot is slightly taller and more rigid than the fluffy huggable Tega robot, which Westlund et al. (2016) used, and we observed that some 3-year-old children were somewhat intimidated by the NAO robot on their first encounter. Such a first impression of the robot could reduce the trust that the child had for the robot, which could negatively affect their willingness to interact with the robot in the short-term, but also in the long-term. To develop a successful first encounter and to build

trust between the child and robot, we designed the following strategy for introducing the robot to 3-year-old children at their preschool.

Pilot studies revealed that some children got anxious when the robot was introduced and then suddenly started to move. To familiarize children prior to their first encounter with the robot, it is therefore advisable to prepare them well. For our study, we sent coloring pages of the robot to the preschools during recruitment and asked the pedagogical assistants to talk a little bit about the robots to the children. About 1 week before the experimental trials, the experimenters introduced the robot in class during their daily "circle time", as this provided a safe and familiar environment with the whole group in which the pedagogical assistants usually introduce new topics or new activities. One experimenter first introduced the robot by telling a story about Robin, the name of our robot, using a makeshift picture book. In this story we explained the similarities and dissimilarities between the robot and children to construct the type of common ground considered to have a positive effect on the learning outcome (Kanda et al., 2004). For example, we told that Robin enjoys dancing and wants to meet new friends, and even though he does not have a mouth and because of that cannot smile, he can smile using his eye LEDs.

After this story, another experimenter entered the room with the robot while it was actively looking at faces to provide an animate feeling. The robot introduced itself with a small story about itself and by performing a dance in which the children were encouraged to participate. The end of the circle time consisted of getting a blanket for the robot so it could "sleep". This introduction was repeated later on the days we conducted the experiment in one-on-one sessions. While by then most children were comfortable interacting with the robot, some were still timid and anxious. To encourage these children to feel comfortable, one of the experiment leaders would sit next to the child during the warm-up phase of the experiment and motivate the child to respond to the robot when necessary until the child was sufficiently comfortable to interact with the robot by herself/himself. We found that the younger 3-year olds required more support from the experimenters than the older 3-year olds (Baxter et al., 2017). Although we are still analyzing the experiments, preliminary findings suggest that our introduction helped children to build trust and common ground with the robot effectively.

Temporal Contingency

Research has shown that it is crucial for children's language development that their communication bids are responded to in a temporally contingent manner (Bornstein et al., 2008; McGillion et al., 2013). This, however, faces a technological challenge. While adults tend to take over turns very rapidly, robots require relatively long processing time to produce a response. Nevertheless, in our first experiment (de Haas et al., 2016), we observed that children were at first surprised by the delayed responses, but quickly adapted to the robot and waited patiently for a response. Perhaps this is because children also require longer than adults to take turns (Garvey and Berninger, 1981) and having framed the robot as a peer children made the delays more plausible or expected. Nevertheless, while a lag in temporal contingency may not harm the interaction with children, it may harm learning. One way to remedy this may be to have the robot start responding by providing a backchannel signal, such as "uhm" to indicate the robot is (still) taking his turn, but requires more time to process (Clark, 1996).

Semantic Contingency

Robots should not only respond to children in a timely fashion, but also in a semantically contingent fashion (i.e., consistent with the child's focus of attention), as this too has a positive effect on children's language acquisition (Bornstein et al., 2008; McGillion et al., 2013). For instance, research has shown that by responding in a semantically contingent manner, either verbally or by following children's gaze, (joint) attention is sustained for a longer duration (Yu and Smith, 2016), allowing children to learn more about a situation. To achieve semantically contingent responses, the robot should be able to understand the child's communication bids, construct joint attention with the child, or at least identify what the child is attending to. Monitoring children's behavior and establishing joint attention are therefore considered crucial for designing a successful robot tutor.

Monitoring Children's Behavior

To understand children's communication bids, as well as to test their pronunciation of the L2, it is important that the robot be equipped with well-functioning automatic speech recognition (ASR). However, the performance of state-of-the-art ASR for children is still suboptimal, especially for preschool-aged children (Fringi et al., 2015; Kennedy et al., 2017). Reasons for this include that children's pronunciation is often flawed and that their speech has a different pitch than adults. Moreover, relatively little research has been carried out in this domain and not much data exist to train ASR on. While it can be expected that the performance of ASR for children will improve in the not too distant future (Liao et al., 2015), until then alternative strategies need to be developed that do not (exclusively) rely on ASR.

In our project, we explore various strategies to achieve this, both based on monitoring non-verbal behaviors of the children and focusing on comprehending rather than producing L2. The first strategy relies on providing children tasks they have to perform in the learning environment, such as placing "a toy cow behind a tree" when teaching spatial language. This, however, requires the visual object recognition on the robot to work well, which is only the case when the scene contains a limited set of distinctively recognizable objects, such as distinctly colored objects (Nguyen et al., 2015). A potential solution explored in our project is to use objects with build-in RFID sensors that can be tracked automatically. The second solution we explore is to use a touch screen tablet that displays scenes the child can manipulate, which not only has the advantage of avoiding the problem of object recognition, but also allows us to control

the robot's responses and vary the scenes in real time. A downside, however, is that it takes away the 3-dimensional physical aspect of embodied cognition that would help the children to better entrench what they learn (Glenberg, 2010). Currently, experiments are underway to investigate the effect of using real vs. virtual objects. These solutions not only aid in understanding the child's communication bids, it also helps in identifying their attention and can thus contribute to establishing joint attention.

Joint Attention and Gestures

Joint attention, where interlocutors attend on the same referent, is a form of social interaction that has been shown to support children's language learning (Tomasello and Farrar, 1986). One way to establish joint attention with a child is to guide their attention to a referent using gestures, such as pointing or iconic gestures. The ability to produce gestures in the real world is potentially one of the main advantages of using physical robots as opposed to virtual agents, who may have a harder time to establish joint attention. However, many robots' physical morphologies do not correspond one-to-one to the human body. Hence, many human gestures cannot be translated directly to robot gestures. For instance, the NAO robot that we use in our research has a hand with three fingers that cannot be controlled independently, so index finger pointing cannot be achieved (see **Figure 1**). Will children still recognize NAO's arm extension as a pointing gesture? And if so, will they be able to identify the object the robot refers to? We are currently running an experiment to investigate how NAO's pointing gestures are perceived, and preliminary findings show that participants have difficulty identifying the referred object on a small tablet screen. Similar issues arise when developing other gestures. One of the other non-verbal behaviors we are using is the coloring of NAO's eye LEDSs to indicate the robot's happiness as a form of positive feedback, since the robot cannot smile with its mouth.

Feedback

Feedback, too, is an interactional feature known to help language learning (Matthews et al., 2007; Ateş-Şen and Küntay, 2015). The question is how should the robot provide feedback, such that it is both pleasant and effective for learning? While adults provide positive feedback explicitly, they usually provide negative feedback implicitly by reformulating children's errors in the correct form. In child-child interactions, however, Long (2006) found that there was a clear advantage in learning from explicit negative feedback (e.g., by saying "no, that's wrong, you need to say 'he ran'") when compared to reformulating feedback (the learner says "he runned" and the teacher reacts with "he ran").

FIGURE 1 | NAO pointing to a block with three fingers. (Note that written, informed consent was obtained from the parents of the child for the publication of this image).

To investigate how children experience feedback from a peer robot, we carried out an experiment among 85 3-year-old Dutch-speaking children at preschools in Netherlands (de Haas et al., 2016, 2017). In this experiment, the children interacted with a NAO robot during which they received a short lesson on how to count from 1 to 4 in English. After a short training phase, in which the children were presented with the four counting words twice in relation to body parts and wooden blocks, they were given instructions by the robot to pick up a given number of blocks. While the instructions were given in their native language, the numbers were uttered in English. In response to the child's ability to achieve the task, the robot provided feedback. The experiment followed a between-subjects design with three conditions: adult-like feedback (explicit positive and implicit negative), peer-like feedback (no positive and explicit negative) and no feedback. We did not find significant differences in learning gain between the conditions, probably because the target words were insufficiently often repeated. However, we explored the way in which the children engaged with the robot after they received feedback and we found that children looked less often at the experimenter in the feedback conditions than in the no feedback condition. Further analyses are carried out to evaluate how the children responded to the various forms of feedback to find out what type of feedback would be most effective for achieving both acceptable and effective tutoring interactions.

Zone of Proximity and Adaptivity

Finally, from a pedagogical point of view it is desirable that the interactions between child and robot be sufficiently challenging and varied so that the child has a target to learn from, but at the same time interactions should not be too difficult, because that may frustrate the child causing it to lose interest in the robot (Charisi et al., 2016). In other words, the robot should remain in Vygotsky's Zone of Proximity that supports an effective learning environment (Vygotsky, 1978). In order to achieve this, the robot should be able to keep track of the children's advancements in language learning and perhaps their emotional states during the tutoring sessions, and adapt to these. While the former can be monitored as discussed previously, it may be possible to detect emotional states known to influence learning (e.g., concentration, confusion, frustration and boredom) using methods from affective computing (D'Mello and Graesser, 2012). Using this type of information, it is possible to adapt the tutoring sessions by either reducing or increasing the number of repetitions, and/or change the subject (Schodde et al., 2017).

CONCLUSION

This perspective article presented some design features that we consider crucial for developing a social robot as an effective second language tutor. We believe the robot is most effective when it is framed as a peer, i.e., as a fellow language learner and playmate, but that is designed to use adult-like interaction strategies to optimize learning efficacy. In order to establish common ground and trust to facilitate long-term interactions, we consider it essential that the robot be introduced with appropriate care on the first encounter. As an example, we outlined our strategy for introducing a robot to preschool children. Interactions between child and robot should be contingent and multimodal, and provide appropriate forms of feedback. We argued that the robot should remain within Vygotsky (1978) Zone of Proximal Development and thus should adapt to the individual level of the child.

We also discussed some technical challenges that need to be solved in order to implement contingent interactions; the most important of which we believe is ASR, which presently does not work well for children's speech. While various technical challenges still remain, we expect that social robots will provide effective digital technologies to support second language development in the years to come.

The present list of design features covers many aspects that need to be considered when developing a tutor robot, but it is not yet comprehensive. One aspect that has not been covered, for instance, concerns the design of robots for children from different cultures, which could require different design choices (Shahid et al., 2014). For example, in some cultures education is more teaching-centered (Hofstede, 1986) and thus designing the tutor as a peer robot may be less effective or acceptable (Tazhigaliyeva et al., 2016). Concluding, this perspective article offers only a first step towards a comprehensive list of design features for tutor robots and additional research is needed to complete and optimize the list.

ETHICS STATEMENT

The Research Ethics Committee of Tilburg School of Humanities approved this study, and the parents of all participating children gave written informed consent in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

PV, MH and EK designed the conceptual aspects of the article; PV, MH, CJ and PB carried out the literature review; PV, EK and MH designed the feedback study; MH, CJ and PB designed the introduction study; MH, CJ and PB carried out the studies; PV and MH wrote the article; CJ, PB and EK revised the article critically.

FUNDING

This work has been supported by the EU H2020 L2TOR project (grant 688014). CJ and PB thank the research trainee program of the Tilburg School of Humanities for their support.

ACKNOWLEDGMENTS

The authors wish to thank all members of the L2TOR project for their support and advice regarding this research. We also thank Kinderopvanggroep Tilburg and all participating daycare centers and preschools for their assistance in this research. Finally, a big thank you to all the children and their parents for participating in our research.

REFERENCES

Ateş-Şen, B. A., and Küntay, A. C. (2015). Children's sensitivity to caregiver cues and the role of adult feedback in the development of referential communication. *The Acquisition of Reference*, eds L. Serratrice and S. E. M. Allen (Amsterdam: John Benjamins), 241–262.

Baxter, P., De Jong, C., Aarts, A., de Haas, M., and Vogt, P. (2017). "The effect of age on engagement in preschoolers' child-robot interactions," in *Companion proceedings of the 12th Annual ACM International Conference on Human-Robot Interaction* (HRI'17), (Vienna, Austria).

Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. E. J., Kopp, S., et al. (2015). "L2TOR-second language tutoring using social robots," in *First Workshop on Educational Robots* (WONDER), (Paris, France).

Bornstein, M. H., Tamis-LeMonda, C. S., Hahn, C. S., and Haynes, O. M. (2008). Maternal responsiveness to young children at three ages: longitudinal analysis of a multidimensional, modular and specific parenting construct. *Dev. Psychol.* 44, 867–874. doi: 10.1037/0012-1649.44.3.867

Charisi, V., Davison, D., Reidsma, D., and Evers, V. (2016). "Children and robots: a preliminary review of methodological approaches in learning settings," in *2nd Workshop on Evaluating Child Robot Interaction - HRI*, (Christchurch, New Zealand).

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot interaction. *Philos. Trans. R. Soc. B Biol. Sci.* 1480, 679–704. doi: 10.1098/rstb.2006.2004

de Haas, M., Baxter, P., de Jong, C., Vogt, P., and Krahmer, E. (2017). "Exploring different types of feedback in preschooler and robot interaction," in *Companion proceedings of the 12th Annual ACM International Conference on Human-Robot Interaction* (HRI'17), (Vienna, Austria).

de Haas, M., Vogt, P., and Krahmer, E. J. (2016). "Enhancing child-robot tutoring interactions with appropriate feedback," in *Proceedings of First Workshop on Long-Term Child-Robot Interaction. IEEE Ro-Man*, (New York, NY).

D'Mello, S., and Graesser, A. (2012). Dynamics of affective states during complex learning. *Learn. Instr.* 22, 145–157. doi: 10.1016/j.learninstruc.2011.10.001

Fridin, M. (2014). Kindergarten social assistive robot: first meeting and ethical issues. *Comput. Hum. Behav.* 30, 262–272. doi: 10.1016/j.chb.2013.09.005

Fringi, E., Lehman, J., and Russell, M. J. (2015). "Evidence of phonological processes in automatic recognition of children's speech," in *16th Annual Conference of the International Speech Communication Association*, (Dresden, Germany), 1621–1624.

Garvey, C., and Berninger, G. (1981). Timing and turn taking in children's conversations. *Discourse Process.* 4, 27–57. doi: 10.1080/01638538109544505

Glenberg, A. M. (2010). Embodiment as a unifying perspective for psychology. *Wiley Interdiscip. Rev. Cogn. Sci.* 4, 586–596. doi: 10.1002/wcs.55

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., and Freynik, S. (2014). Technologies for foreign language learning: a review of technology types and their effectiveness. *Comput. Assist. Lang. Learn.* 27, 70–105. doi: 10.1080/09588221.2012.700315

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

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, 517–527. doi: 10.1177/0018720811417254

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

Hofstede, G. (1986). Cultural differences in teaching and learning. *Int. J. Intercult. Relat.* 10, 301–320. doi: 10.1016/0147-1767(86)90015-5

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*, (New York, NY), 83–90.

Hsin, C.-T., Li, M.-C., and Tsai, C.-C. (2014). The influence of young children's use of technology on their learning: A. *Educ. Technol. Soc.* 17, 85–99. Available online at: https://eric.ed.gov/?id=EJ1045554

Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: a field trial. *Hum. Comput. Interact.* 19, 61–84. doi: 10.1207/s15327051hci1901&2_4

Kennedy, J., Baxter, P., Senft, E., and Belpaeme, T. (2015). "Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions," in *International Conference on Social Robotics*, eds A. Tapus, E. André, J.-C. Martin, F. Ferland and M. Ammi (New York, NY: Springer International Publishing), 327–336.

Kennedy, J., Baxter, P., Senft, E., and Belpaeme, T. (2016). "Social robot tutoring for child second language learning," in *Proceedings of the 11th Annual ACM/IEEE International Conference on Human-Robot Interaction* (HRI'16), (Christchurch, New Zealand), 231–238.

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 12th Annual ACM International Conference on Human-Robot Interaction* (HRI'17), (Vienna, Austria).

Leyzberg, D., Spaulding, S., Toneva, M., and Scassellati, B. (2012). "The physical presence of a robot tutor increases cognitive learning gains," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, (Sapporo, Japan), 1882–1887.

Liao, H., Pundak, G., Siohan, O., Carroll, M. K., Coccaro, N., Jiang, Q. M., et al. (2015). "Large vocabulary automatic speech recognition for children," in *Interspeech*, (Dresden, Germany), 1611–1615.

Long, M. H. Ed. (2006). "Recasts in SLA: the story so far," in *Problems in SLA. Second Language Acquisition Research Series*, (Mahwah, NJ: Lawrence Erlbaum Associates), 75–116.

Lund, H. H. (2003). "Adaptive robotics in the entertainment industry," in *Proceedings of IEEE International Symposium on Computational Intelligence in Robotics and Automation (cira2003)*, (Vol. 2), (Kobe, Japan), 595–602.

Mashburn, A. J., Justice, L. M., Downer, J. T., and Pianta, R. C. (2009). Peer effects on children's language achievement during pre-kindergarten,. *Child Dev.* 80, 686–702. doi: 10.1111/j.1467-8624.2009.01291.x

Matthews, D., Lieven, E., and Tomasello, M. (2007). How toddlers and preschoolers learn to uniquely identify referents for others: a training study. *Child Dev.* 6, 1744–1759. doi: 10.1111/j.1467-8624.2007.01098.x

McGillion, M., Herbert, J., Pine, J., Keren-Portnoy, T., Vihman, M., and Matthews, D. (2013). Supporting early vocabulary development: what sort of responsiveness matters? *IEEE Trans. Auton. Ment. Dev.* 5, 240–248. doi: 10.1109/tamd.2013.2275949

Nguyen, T. L., Boukezzoula, R., Coquin, D., Benoit, E., and Perrin, S. (2015). "Interaction between humans, NAO robot and multiple cameras for colored objects recognition using information fusion," in *8th International Conference on Human System Interaction* (HSI), (Warsaw, Poland), 322–328.

Schodde, T., Bergmann, K., and Kopp, S. (2017). "Adaptive robot language tutoring based on bayesian knowledge tracing and predictive decision-making," in *Proceedings of the 12th ACM/IEEE International Conference on Human-Robot Interaction* (HRI 2017), (Vienna, Austria).

Shahid, S., Krahmer, E., and Swerts, M. (2014). "Child–robot interaction across cultures: how does playing a game with a social robot compare to playing a game alone or with a friend? *Comput. Human Behav.* 40, 86–100. doi: 10.1016/j.chb.2014.07.043

Tanaka, F., and Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: field experiments in a classroom for vocabulary learning. *J. Hum. Robot Interact.* 1, 78–95. doi: 10.5898/jhri.1.1.tanaka

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," in *International Conference on Social Robotics*, eds A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs and H. He (New York, NY: Springer International Publishing), 327–336.

Tomasello, M., and Farrar, M. J. (1986). Joint attention and early language. *Child Dev.* 57, 1454–1463. doi: 10.2307/1130423

Vygotsky, L. (1978). *Mind in Society*. Harvard: Harvard University Press.

Westlund, J. M. K., 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 *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, eds S. Y. Okita, T. Shibata and B. Mutlu (Washington, DC: IEEE), 688–693.

Yu, C., and Smith, L. B. (2016). The social origins of sustained attention in one-year-old human infants. *Curr. Biol.* 26, 1235–1240. doi: 10.1016/j.cub.2016.03.026

Conflict of Interest Statement: 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.

2017-03-06

The Effect of Age on Engagement in Preschoolers' Child-Robot Interactions

Baxter, P

10.1145/3029798.3038391

The Effect of Age on Engagement in Preschoolers' Child-Robot Interactions

Peta Baxter[1], Chiara de Jong[1], Rian Aarts[2], Mirjam de Haas[1], Paul Vogt[1]
[1] Tilburg center for Cognition and Communication, [2] Department of Culture Studies, Tilburg School of Humanities
Tilburg University
Tilburg, The Netherlands
p.a.baxter@tilburguniversity.edu, p.a.vogt@tilburguniversity.edu

ABSTRACT
In this study, we investigate the effect of age on preschoolers' engagement - as measured by gaze direction - during a first-time interaction with a social robot. The results revealed significant differences in gaze patterns. Specifically, younger children were more easily distracted, and looked at the robot for a shorter duration and briefer periods of gaze. Moreover, they showed a higher level of reliance on the experimenters. The results have implications for the design of young preschoolers child-robot interactions and specifically for the ways in which the first introductory interactions should occur.

Keywords
Child-robot interaction; engagement; social robots; gaze

1. INTRODUCTION
In recent years, there has been an increasing effort to develop and integrate robots as peer-tutors in (pre)schools, for example for the purpose of teaching foreign languages [1]. While most studies have focused on school-aged children, current research is also targeting preschool children, who have high learning flexibility [2]. However, preschool-aged children (2 to 4 years old) undergo major cognitive, emotional and social developments, such as expanding their social competence [3, 4], which must be accounted for in such studies. Whereas older children may have little difficulty engaging in an interaction with a robot, younger children may be more reliant on their caregivers or show less engagement in the interaction, as they are less socially competent. Children between the ages of 3 and 4 show substantial differences in emotional competence, which predicts later social competence [4]. Therefore, we expect that child-robot interactions at those ages will also present some age-related variation. Clarifying these potential age differences is essential as, in order to be efficient, interactive scenarios with robots must be tailored to the diverging needs of children. In the current study, we sought to determine whether there are age-related differences in first-time interactions with a peer-tutor robot of children who have just turned 3 and children who are almost 4 years old. Specifically, we hypothesized that younger children may experience more difficulty engaging with a robot and may rely more heavily on adults if it is their first one-on-one interaction with a robot.

HRI'17 Companion, March 6–9, 2017, Vienna, Austria.
ACM ISBN 978-1-4503-4885-0/17/03
DOI: http://dx.doi.org/10.1145/3029798.3038391

Since previous research has shown that gaze behavior is a good indicator of engagement, especially in human-agent interaction [5], we measured preschoolers' engagement by means of their gaze direction.

2. METHODS
Thirty-two children recruited at preschools in the Netherlands participated in this study (18 female, M = 41.47 months, SD = 4.74) of which 17 were in the young age group (M = 37.35, SD = 2.06) and 15 were in the old age group (M = 46.13, SD = 0.99). Prior to a one-on-one interaction with the NAO robot, the children took part in a group introduction to familiarize them with the robot. Two experimenters were present during the one-on-one interaction. They kept in the background, only intervening when children required it. The full interaction was filmed, and consisted of an introductory phase followed by a short tutoring session for English as a second language revolving around counting blocks. For this study, we only considered the introductory phase, since we were interested in the initial response to the robot. During this phase (Mduration = 5.9 minutes, SD = 1.09) the robot introduced itself and initiated a conversation that encouraged an exchange of personal information. Additionally, a few simple counting tasks revolving around the blocks and the robot's body parts were included. Children were filmed from two viewpoints to account for erratic movements. Gaze behavior (frequency and duration) was analyzed by manually coding the children's gaze towards the robot, the experimenter(s), the blocks, themselves and elsewhere (Cohen's Kappa = .82). Glances, i.e. gaze shorter than 1 second, were not considered to be an actual gaze pattern and were therefore added to the nearest annotation.

3. RESULTS
To explore the differences in gaze behaviors within each group, we conducted Greenhouse-Geisser corrected repeated-measures ANOVAs (see Figure 1 and 2 for visual representations).

For the younger children, this revealed significant differences in gaze frequency and proportion of time in a certain gaze direction, respectively $F(2.49, 39.82) = 39.89$, $p < .001$ and $F(2.06, 30.47) = 84.79$, $p < .001$. Specifically, younger children looked at the experimenters more frequently and for a longer proportion of time than elsewhere (respectively, $p = .007$; $p = .012$), themselves (respectively, $p = .001$; $p = .001$) and more frequently at the blocks ($p = .007$). Overall though, younger children also looked at the robot more frequently and for a longer proportion of time than at the experimenters ($p < .01$), the blocks, elsewhere, or themselves, all $p < .001$. For the older children, we found significant differences with regard to gaze frequency and proportion of time in a certain gaze direction, respectively $F(2.35, 32.84) = 21.77$, $p < .001$ and $F(1.37, 19.20) = 109.43$, $p < .001$. Specifically, they looked at the experimenters more frequently than elsewhere ($p = .038$) and at the blocks for a longer proportion

of time than elsewhere (p = .026). Just like the younger children, they looked at the robot more frequently and for a longer proportion of time than the experimenters (p < .01), the blocks, elsewhere, or themselves, all p < .001.

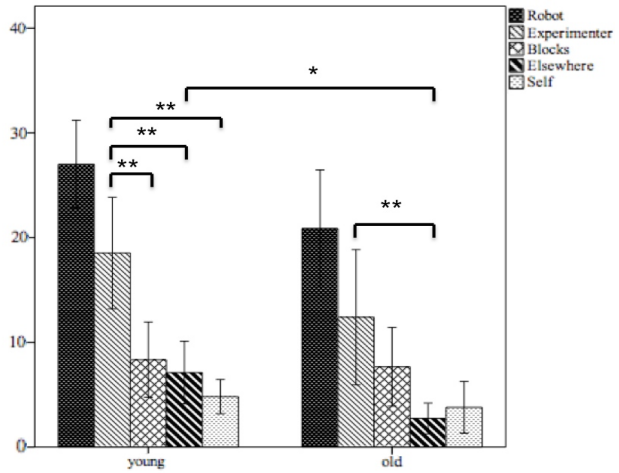


Figure 1. Mean number of occurrences for each gaze direction

*Note. *p<.05, **p<.01. Robot gaze differed significantly from all other gaze directions, p < .01*

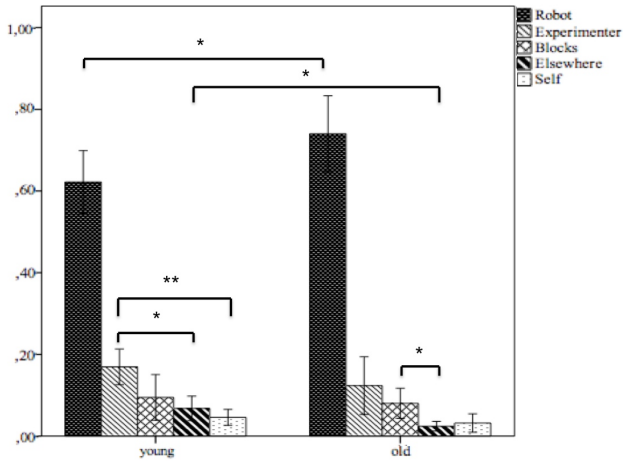


Figure 2. Proportion of time spent on each gaze direction

*Note. *p<.05, **p<.01. Robot gaze differed significantly from all other gaze directions, p < .01*

To analyze the effect of age, t-tests were conducted, both for gaze frequency and proportion of time. Older children looked at the robot for a larger proportion of time (M = .74, SD = .17) than younger children (M = .62, SD = .15), $t(30)$ = -2.11, p = .043. The average time (in seconds) per gaze on the robot was higher for older children (M = 14.97, SD = 8.60) than for younger children (M = 9.74, SD = 5.73), $t(30)$ = -2.05, p = .049. Additionally, younger children looked elsewhere for a larger proportion of time (M = .07, SD = .06) and more frequently (M = 7.12, SD = 5.81) than older children (respectively M = .02, SD = .02, M = 2.73, SD = 2.66), respectively $t(30)$ = 2.74, p = .010 and $t(30)$ = 2.68, p = .012.

4. DISCUSSION AND CONCLUSION

The current study sought to determine the effect of age on preschoolers' engagement during first time one-on-one child-robot interactions. The results indicate that, between the ages of 3 and 4, age differences as small as 10 months lead to diverging engagement behaviors towards a robot. In our experiment, both younger and older children looked at the robot more often and for a longer proportion of time than anywhere else, illustrating the overall interest in the robot. However, younger children spent less time - overall as well as per gaze - looking at the robot than older children. They also looked elsewhere more often and for a longer proportion of time. This suggests that while younger children do show interest in the robot and are engaged with it, they might be less able to sustain direct attention towards it than older children.

We postulate that these results are caused by the fact that younger children are more easily distracted by their surrounding and have more trouble focusing on a task for an extended period of time, unlike the older children, who were mainly focused on the robot. In addition, given that younger children looked at the experimenters more often and for a larger proportion of time than anywhere else (other than the robot), we hypothesize that they need additional support, reassurance and feedback in their first interaction with a robot. For instance, it was relatively common for the younger children to look at the experimenters after they had answered one of the robot's questions. Further analyses of the experimenters' interventions and children's requests for help should contribute to verifying this hypothesis. The results of the current study have implications for the design of (first-time) interactions between preschoolers and social robots, with special attention required to providing suitable support for the youngest children.

5. ACKNOWLEDGMENTS

This work has been supported by the EU H2020 L2TOR project (grant 688014). The authors would like to thank the research trainee program of the Tilburg School of Humanities for their support. We also thank Kinderopvanggroep Tilburg and all preschools for participating in this research.

6. REFERENCES

[1] Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Hum.-Comput Interact.*, 19, 61-84.

[2] Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. E., Kopp, S., ... and Pandey, A. K. 2011. L2TOR-Second Language Tutoring using Social Robots. In *Proceedings of the 1st Int. Workshop on Educ. Robots*. Springer

[3] Denham, S. A., and Couchoud, E. A. 1990. Young preschoolers' understanding of emotions. *Child Study J.*, 20, 171-192.

[4] Denham, S. A., Blair, K. A., DeMulder, E., Levitas, J., Sawyer, K., Auerbach–Major, S., and Queenan, P. 2003. Preschool emotional competence: Pathway to social competence?. *Child Dev.*, 74, 238-256.

[5] Nakano, Y. I., & Ishii, R. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th Int. conference on Intelligent User Interfaces* (Hong-Kong, China, February 07 - 10, 2010). ACM, New York, NY, 139-148.

3 2016 Publications

- Mirjam de Haas, Paul Vogt and Emiel Krahmer (2016) Enhancing child-robot tutoring interactions with appropriate feedback. In *Proceedings of the Long-Term Child-Robot Interaction Workshop, at RO-MAN 2016.*

- Paul Baxter, James Kennedy, Emily Ashurst and Tony Belpaeme (2016) The Effect of Repeating Tasks on Performance Levels in Mediated Child-Robot Interactions. In *Proceedings of the Robots 4 Learning Workshop, at RO-MAN 2016.*

- James Kennedy, Séverin Lemaignan and Tony Belpaeme (2016) The Cautious Attitude of Teachers Towards Social Robots in Schools. In *Proceedings of the Robots 4 Learning Workshop, at RO-MAN 2016.*

- Séverin Lemaignan, James Kennedy, Paul Baxter and Tony Belpaeme (2016) Towards "Machine-Learnable" Child-Robot Interactions: the PInSoRo Dataset. In *Proceedings of the Long-Term Child-Robot Interaction Workshop, at RO-MAN 2016.*

- Paul Baxter, and Tony Belpaeme (2016) A Cautionary Note on Personality (Extroversion) Assessments in Child-Robot Interaction Studies. In *Proceedings of the 2nd Workshop on Evaluating Child-Robot Interaction, at HRI'16.*

- Paul Baxter, James Kennedy, Emmanuel Senft, Séverin Lemaignan, and Tony Belpaeme (2016) From Characterising Three Years of HRI to Methodology and Reporting Recommendations. In *Proceedings of the 11th Annual ACM/IEEE International Conference on Human-Robot Interaction (alt.HRI'16).* Pages 391-398.

- James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme (2016) Heart vs Hard Drive: Children Learn More From a Human Tutor Than a Social Robot. In *Proceedings of the 11th Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts (HRI'16).* Pages 451-452.

- James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme (2016) Social Robot Tutoring for Child Second Language Learning. In *Proceedings of the 11th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'16).* Pages 231-238.

- Mirjam de Haas, Paul Vogt and Emiel Krahmer (2016) Taal leren met behulp van een sociale robot. In *DIXIT (Tijdschrift van de Nederlandse Organisatie voor Taal- en Spraaktechnologie NOTAS).* *

*No pdf available

Enhancing child-robot tutoring interactions with appropriate feedback

Mirjam de Haas[1], Paul Vogt[1] and Emiel Krahmer[1]

Abstract— **The use of social robots for teaching children a second language is a promising field. This paper describes an ongoing experiment in which we explore how children engage with a robot after receiving feedback in a tutoring session. We created three scenarios in which the robot performed peer-like, adult-like feedback or was withholding feedback. A group of 85 preschool children are investigated. We will compare how the children experience interaction with the robot and their responses to the different types of the robot's feedback. The purpose is to explore the possibilities of peer interaction between robot and child for long-term tutoring.**

I. Introduction

Robots are starting to enter the classroom and more and more children are interacting with robots. Studies have shown that robots can have effective interactions with children with regards to educational settings [1]. Children are less stressed when interacting with robots and are more readily engaged with tasks that are otherwise considered boring. Robots are able to teach children different school subjects, one of these is second language [2], [3]. This can be tutoring a popular second language such as English or tutoring the official school language to children from immigrant families. Teaching immigrant children the school language is crucial in early stages of development, as later educational success builds on that [4].

In educational settings, children are expected to maintain long-term relationships with a tutoring robot. To achieve successful long-term relationships, natural and varied interactions between the robot and children are crucial. One of the challenges is to keep the child interested. For establishing long-term relations, a robot should engage users over extended periods of time and this can be achieved through an understanding of interactions between peers [5]. In most educational settings, the robot acts as a teacher (see for example [6]). However, younger children prefer robots to behave as peers and, within language, they prefer a tutoring style [7]. A child would perceive a peer tutor as a friend with more advanced language skills, would interact with the peer tutor as a friend and would receive feedback from the peer tutor as a friend. Observations of peer interactions between children [9] showed that children provide less feedback than adults and they produce different feedback when their peers make mistakes. However, in interactions between native and non-native children's interactions, non-native children receive significantly more feedback than the native children. Therefore, to create a robot behaving as a peer and a tutor we expect that children respond to the robot as a peer and the robot would also respond to the children as a peer. Moreover, a robot that gives appropriate feedback is expected to support the child's language development best. Question is: How should a robot provide feedback to make the interaction both pleasant and educational?

Klugel and DeNisi mention that no feedback is sometimes better [10]. In a meta-analysis of 131 studies they found one third of these studies show that there are negative effects of feedback compared to no feedback at all. However, they did not investigate the type of feedback, only the amount. Negative feedback might have more impact on learning efficacy, although positive feedback can give some reassurance to the learner [11]. Older people showed a higher user compliance and performance when a robot gave feedback during their workout [12], [13]. Moreover, robots giving positive feedback is widely used within therapy when children have Autism Spectrum Disorder (ASD) [14]. In addition, children with ASD tend to be more motivated when the robot gives a reward after a correct behavior. When a robot reacts to our actions it makes us more confident in the robot's behavior.

We want to investigate whether these results can be extended to typically developing preschool children learning a second language. In child-child interactions, Long [15] found that there was a clear advantage in learning for explicit feedback (e.g. by saying "no, that's wrong, you need past tense") when compared to recasting feedback (the learner says "he runned" and the teacher reacts with "he ran"). The explicitness of the feedback is also an important determinant of children's responses to feedback. In a free-play situation where four-year old children could play, observations revealed that children responded much more often to specific questions than for implicit nonverbal feedback, or implicit verbal feedback [16], [17].

Mazzoni [18], explored feedback of a humanoid robot in language learning of young children. Children were asked to play with either the robot or another child, and work together to understand the meaning of an English word. The robot did not give explicit feedback, but it introduced a doubt (for example, "ahh, your suggestion is interesting … but are we sure that it is correct?"). If children did not respond, the robot would ask them for suggestions. The children showed in both conditions (one in interaction with the robot and one with another child) improvement in their Engels vocabulary. The authors, however, did not provide information on how children reacted on the robot and whether the children considered the robot as a peer or else.

*Research supported by European Union's Horizon 2020 and innovation programme under grant No 688014.

[1]Tilburg center for Cognition and Communication, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands
{mirjam.dehaas, p.a.vogt, e.j.krahmer}@tilburguniversity.edu

The objective of this study is to answer the question whether the type of feedback (more explicit such as peers or implicit such as adults) that the robot gives to children will influence their engagement with the robot, and compare this with a robot that gives no feedback at all.

Mashburn et al. [20] found that peer interactions have a positive effect on language development. However, children with relative poor language abilities benefited less from peer interactions, because they had less opportunity to engage with other children. One-to-one interactions in which a robot provides opportunities to engage more, might be less intimidating than an actual peer and can have a positive effect on the children's language abilities. Other than adults, children are more focused on constructing their own personal meaning, and, therefore, use less negotiation techniques that focus on their peers' understanding. Furthermore, Mackey et al. [9] investigated patterns of negotiation in child interactions and found that the children use three different forms of negotiation: clarification (what do you mean?), confirmation (do you mean this?) and comprehension (do you understand?).

This research is part of the L2TOR project, which focuses, among others, on teaching native Dutch children English as a second language, and on teaching Dutch to native speakers of Turkish living in the Netherlands [19]. The general idea is that the robot will support all children in both their native language and the second language.

The remainder of this paper describes an experiment that investigates the influence of providing peer-like or adult-like feedback by the robot, aged 3 to 4 years on the child's engagement with the robot. The experiment was carried out in various preschools in the Netherlands. We created a scenario in which the children interact with a humanoid robot either giving one of the feedback types or withholding feedback and study the effect of the robot on the children's engagement in the activity and their relationship with the robot.

II. Experimental Design

This experimental design will describe an experiment in which a robot will teach Dutch speaking children English. We will explore the children's reactions on the robot's feedback.

A. Participants

Approximately 85 preschool children of 3 to 4 years old will take part in this experiment. These children attend a preschool in Tilburg and are normally instructed in Dutch. For all children the parents sign an informed consent form.

B. Task

The task is a collaborative game with blocks. The robot uses the blocks to teach the children to count from 1 to 4 in English. During the interaction the robot instructs all children in Dutch and only names the different numbers in English. Each child sits on the ground in front of the robot that is approximately 40 cm from the participant (see Fig. 1). The experimenter explains the children that the robot is going to teach the children some words in English. The duration of the interaction is around 10 to 15 minutes, depending on how much feedback the children need. Prior to the experiment the children practice counting in Dutch together with the experimenter and the blocks and their knowledge of the English counting words will be tested before and after the experiment. The children were not given any feedback during the pre and posttest.

Fig. 1. Experimental setup

C. Robot

The robot used during this experiment is the Nao robot, which is a small humanoid robot produced by Aldebaran. This robot has already been used in many studies with children. The advantage of using Nao is that this robot can use gestures to explain the children the words. Children are more engaged when the speech is accompanied by gestures, their joint attention increases the interactions are longer and they look more at the robot during its turn [21], [14]. The robot points and gazes at the blocks that are used in this experiment. Moreover, it gazes at the children during interactions. The children were already introduced to the robot prior to the experiment and were explained how the robot shows emotions and they were familiar with the behaviors of the robot. Furthermore, the robot speaks with a synthesized Dutch and English voice. Most of the sentences will be in Dutch; only the target words for the children are in English. While we plan to use automatic speech recognition in the near future, we use the Wizard of Oz method [22], because of the imperfections in the automatic speech recognition of child speech. This way, it appears for the children as if the robot is responding on their questions and actions.

D. Experimental Conditions

In this study we want to test the impact of different types of feedback. All conditions are tested with a between subject design. The children are randomly assigned in one of the feedback conditions. The behavior and movements of the robot remain identical between all conditions, except for the robot's feedback.

We use two types of feedback; the first one uses explicit feedback that children often use during peer interactions and the second one uses implicit recasting feedback that adults most often use while interacting with children and compare these to a condition without feedback. Prior research has

shown that children react more often to explicit specific questions [16] and we, therefore, included explicit egocentric feedback in the peer-feedback condition. The other type of feedback is based on how adults respond to children and how they interact with them. Adults use recasting feedback and tend to praise the children for their work. This adult-feedback condition contains of implicit (recasting) feedback and giving praise to the child whenever they did something correct.

In the examples below, the text said in English is indicated in Italics, the rest of the text is said in Dutch.

1) No Feedback condition

This condition is the baseline condition for this experiment, wherein the robot only serves as a language instructor and playmate for this game. The robot does not explicitly motivate the child by giving feedback. All motivations come from the other instructions and the child's own intrinsic motivation. When the experimenter notices that the child does something completely wrong with the result that interaction does not continue, she corrects the mistake of the child after the interaction with the robot.

Example of no feedback after correct and incorrect child response:
Robot: "Can you show me *three* blocks?"
Learner: shows robot three blocks.
Robot: "Put all the blocks back. Can you show me *two* blocks?"

2) Peer-Feedback condition

In this condition the scenario sequence is the same as in the no feedback condition, with an addition that the robot gives explicit feedback whenever the child does something wrong. The verbal feedback changes every time, only the non-verbal feedback stays the same during the task itself.

Example of feedback after correct child response:
Robot: "Can you show me *three* blocks?"
Learner: shows robot three blocks.
Robot: "Put all the blocks back. Can you show me *two* blocks?"

Example of feedback after incorrect child response:
Robot: "Can you show me *three* blocks?"
Learner: shows robot two blocks.
Robot: "That's wrong! You should take three blocks."

3) Adult-Feedback condition

In this condition the scenario sequence is the same as in the other two conditions, except that the robot gives feedback when the child responds either correctly or incorrectly. When the child responds correctly, the robot gives positive feedback both verbally and non-verbally by showing the child that it is happy by blinking its eyes in different colors. When the child makes a mistake, implicit negative feedback is provided, which is less strong as in the peer-feedback condition.

Example of feedback after correct child response:
Robot: "Can you show me *three* blocks?"
Learner: shows robot three blocks.
Robot: "Well done! *Three* means three in English."

Example of feedback after incorrect child response:
Robot: "Can you show me *three* blocks?"
Learner: shows robot two blocks.
Robot: "*Three* means three, you should take three blocks."
Learner: shows robot three blocks
Robot: "Well done! *Three* means three in English."

E. Hypotheses

The main purpose of our experiment is to investigate how the children are engaged with the robot in all conditions, while the effectiveness of the language tutoring is of secondary importance in this experiment. We therefore have the following three hypotheses:

H1. We expect that the robot that gives feedback will engage the children more than the robot that gives no feedback. Mackey explored feedback with children and also found these results, although they did not test this with a robot, we still expect this will be true for the robot and a child [9].
H2. We expect that the children will be more motivated to continue when the robot gives positive feedback.
H3. We expect that the children will learn more target words in the peer-feedback condition due to the explicit negative feedback.

F. Evaluations

The experiments, which are concluded at the moment of writing this paper, have been recorded on video. These recordings will be analyzed for the child's engagement with the robot using a coding scheme adapted from [23]. In particular, we will measure children's reaction to having feedback or not in a perception study. To this aim, short video fragments, displaying children's responses to the feedback of the robot or the absence thereof, will be shown in random order to naive observers. These observers are asked to indicate for each snippet whether the child displays positive or negative emotions, which will indicate how children are engaged with the robot after a certain type of feedback.

Second, we will measure the proportion of time children are engaged with the robot. For this, we will adopt the coding scheme of Mastin and Vogt [24] to assess the amount of time children are engaged with the robot and whether the engagement concerns episodes of joint attention or not.

Finally, we will measure whether there is any learning effect from interacting with the robot. To this aim, we will carry out a short pre-test and a short post-test to test the children's ability to count from 1 to 4 in Dutch and in English.

III. CONCLUSION

This paper described an experiment in which a robot is used to teach children a second language. The experiment described explores how children react to feedback of the robot. The experiment has taken place in June and is

concluded at the moment of writing, so we expect to present some preliminarily results during the workshop in August.

REFERENCES

[1] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. Mcowan, "Multimodal Affect Modeling and Recognition for Empathic Robot Companions," *Int. J. Humanoid Robot.*, vol. 10, no. 01, p. 1350010, 2013.

[2] J. Kennedy, P. Baxter, and T. Belpaeme, "The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning," *Proc. ACM/IEEE Int. Conf. Human-Robot Interact.*, no. January 2016, pp. 67–74, 2015.

[3] M. Fridin, "Storytelling by a kindergarten social assistive robot: A tool for constructive learning in preschool education," *Comput. Educ.*, vol. 70, pp. 53–64, 2014.

[4] P. P. M. Leseman, "Effects of Quantity and Quality of Home Proximal Processes on Dutch , Surinamese – Dutch and Turkish – Dutch Pre - schoolers ' Cognitive Development," vol. 38, no. January 1998, pp. 19–38, 1999.

[5] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz, and J. Wang, "Designing Robots for Long-Term Social Interaction," in *International Conference on Intelligent Robots and Systems*, 2005, pp. 2199–2204.

[6] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive Robots as Social Partners and Peer Tutors for children: A field trial," *Human-computer Interact.*, vol. 19, pp. 61–84, 2004.

[7] M. Saerbeck, T. Schut, C. Bartneck, and M. D. Janse, "Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor," in *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, 2010, pp. 1613–1622.

[8] N. Shin and S. Kim, "Learning about, from, and with robots: Students' perspectives," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2007, pp. 1040–1045.

[9] A. Mackey and J. Leeman, "Interactional input and the incorporation of feedback an exploration of NS-NNS and NNSNNS adult and child dyads.," *Lang. Learn.*, vol. 53, no. 1, pp. 35–66, 2003.

[10] A. N. Kluger and A. DeNisi, "Feedback Interventions: Toward the Understanding of a Double-Edged Sword.," *Am. Psychol. Soc.*, vol. 7, no. 3, pp. 67–72, 1998.

[11] S. Y. Okita and D. L. Schwartz, "Learning by Teaching Human Pupils and Teachable Agents: The Importance of Recursive Feedback," *J. Learn. Sci.*, vol. 22, no. 3, pp. 375–412, 2013.

[12] J. Fasola and M. J. Mataric, "Robot motivator: Improving user performance on a physical/mental task," in *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, 2009, pp. 295–296.

[13] B. J. Fasola and M. J. Mataric, "Using Socially Assistive Human – Robot Interaction to Motivate Physical Exercise for Older Adults," in *Proceedings of the IEEE*, 2012, vol. 100, no. 8, pp. 2512–2526.

[14] E. I. Barakova, P. Bajracharya, M. Willemsen, T. Lourens, and B. Huskens, "Long-term LEGO therapy with humanoid robot for children with ASD," *Expert Syst.*, vol. 32, no. 6, pp. 698–709, 2015.

[15] Long. M. H., "Recasts in SLA: The story so far," in *Problems in SLA. Second Language Acquisition Research Series.*, Mahwah, NJ: Lawrence Erlbaum Associates., 2006, pp. 75–116.

[16] C. L. Petersen, F. W. Danner, and J. H. Flavell, "Developmental Changes in Children ' s Response to Three Indications of Communicative Failure," *Soc. Res. Child Dev.*, vol. 43, no. 4, pp. 1463–1468, 1972.

[17] D. Spilton and L. C. Lee, "Some Determinants of Effective Communication in Four-Year-Olds," *Soc. Res. Child Dev.*, vol. 48, no. 3, pp. 968–977, 1977.

[18] E. Mazzoni and M. Benvenuti, "A Robot-Partner for Preschool Children Learning English Using Socio-Cognitive Conflict," *Educ. Technol. Soc.*, vol. 18, no. 4, pp. 474–485, 2015.

[19] J. Kennedy, P. Baxter, E. Senft, and T. Belpaeme, "Social Robot Tutoring for Child Second Language Learning," 2016, pp. 231–238.

[20] A. J. Mashburn, L. M. Justice, J. T. Downer, and R. C. Pianta, "Peer effects on children's language achievement during pre-kindergarten," *Child Dev*, vol. 80, no. 3, pp. 686–702, 2009.

[21] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artif. Intell.*, vol. 166, pp. 140–164, 2005.

[22] L. D. Riek, "Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines," *J. Human-Robot Interact.*, vol. 1, no. 1, pp. 119–136, 2012.

[23] S. Shahid, E. Krahmer, and M. Swerts, "Child-robot interaction across cultures: How does playing a game with a social robot compare to playing a game alone or with a friend?," *Comput. Human Behav.*, vol. 40, pp. 86–100, 2014.

[24] J. D. Mastin and P. Vogt, "Infant engagement and early vocabulary development: a naturalistic observation study of Mozambican infants from 1;1 to 2;1," *J. Child Lang.*, pp. 1–30, 2016.

The Effect of Repeating Tasks on Performance Levels in Mediated Child-Robot Interactions

Paul Baxter
Lincoln Centre for Autonomous Systems
School of Computer Science
University of Lincoln (U.K.)
Email: pbaxter@lincoln.ac.uk

James Kennedy, Emily Ashurst, Tony Belpaeme
Centre for Robotics and Neural Systems
The Cognition Institute
Plymouth University (U.K.)
Email: first.last@plymouth.ac.uk

***Abstract*—That "practice makes perfect" is a powerful heuristic for improving performance through repetition. This is widely used in educational contexts, and as such it provides a potentially useful feature for application to child-robot educational interactions. While this effect may intuitively appear to be present, we here describe data to provide evidence in support of this supposition. Conducting a descriptive analysis of data from a wider study, we specifically examine the effect on child performance of repeating a previously performed collaborative task with a peer robot (i.e. not an expert agent), if initial performance is low. The results generally indicate a positive effect on performance through repetition, and a number of other correlation effects that highlight the role of individual differences. This outcome provides evidence for the variable utility of repetition between individuals, but also indicates that this is driven by the individual, which can nevertheless result in performance improvements even in the context of peer-peer interactions with relatively sparse feedback.**

I. Introduction

The research-oriented application of social robots to educational contexts (in particular for children) has been rapidly increasing in recent years. The applications have spanned schools [1], healthcare [2], and extracurricular activity scenarios [3], and covered a wide range of subjects and skills, from nutrition [4] to handwriting [5].

Such work frequently attempts to bootstrap from contemporary learning theories. While methods such as learning by rote (effectively memorisation without explicit emphasis on understanding) were formerly stardard educational practice (based partly on behaviourist ideas [6]), more recently constructivist and related approaches have come to the fore [7] even though in practice the behaviourist approach frequently remains in place (the use of reinforcers and testing to name but two). Robotics applications have thus in principle typically followed this latter approach, emphasising concepts such as collaboration [8], social partership [9], guided discovery [10], and others.

Typically, given the as yet novel nature of such child-robot interaction work, these studies generally take a relatively high level perspective, emphasising metrics such as user preference/opinion and/or overall learning effects. However, as such applications mature, it will become necessary to perform more fine-grained analyses in order to establish the conditions under which children may maximise their learning with a robot, and the features (behavioural, morphological, etc) of the robot that best facilitate this learning process. The study described in this paper is presented in this context.

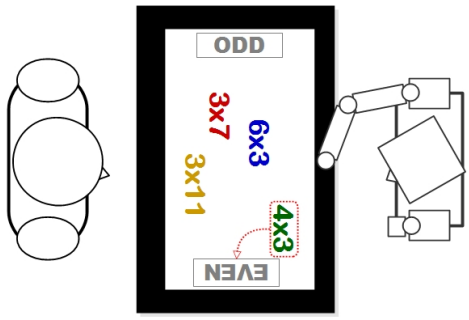


Fig. 1. Child (left) and robot (right) playing a collaborative maths sorting game (categorising the result of the multiplications as either odd or even in this example) on a large touchscreen located between them. Images on screen and dashed sample image path shown for illustration only; not to scale.

In this paper, we describe data that demonstrates the effect on child performance of repeating a collaborative task with a robot, if the performance of an initial attempt is low. Overall, the results do suggest that there is some benefit conferred by such repetition. However, one strong theme that further emerges from this analysis is the high variability of the effect, which indicates the importance of individual differences. First the study is described, highlighting the embedded nature of the experiment in a classroom and the collaborative nature of the task (section II), followed by results that explore overall phenomena and effects at an individual level (section III), before further discussing these effects at the end of the paper (section IV).

II. Study

The aim of the study presented in this paper is to assess the effect of repeating a task with a robot in which initial performance was low. To achieve this, we analyse a sub-set of data obtained from a larger study, which sought to assess the impact of embedded (i.e. present in the classroom itself with no experimenters present) personalised robot peers in a classroom on learning [11]. Using a two-condition setup (personalised intervention condition and non-personalised control condition),

results indicate that personalisation supports additional learning, particularly in novel subjects [11], and that teachers will take advantage of a robot in their classroom for wider moivational purposes than just the task to be performed with the robot [12]. Since performance-based repetition was only present in the intervention condition, this is the data that we analyse below.

Experimental hypotheses are not ventured for the present paper due to its placement within the wider study. As such, we provide a descriptive analysis of the data obtained with respect to the effect of repetition, as a means of providing initial indications of effects that can be subsequently taken up in further studies in their own right.

A. Ethics

Approval for conducting this study was granted by the Plymouth University Faculty of Science and Technology Human Ethics Committee, as part of a thematic programme of research involving the robot and touchscreen setup, and children in local schools. An opt-out informed consent was obtained in writing from the parents/guardians of all participating children. It was made clear to all children that they could withdraw if and when they wished to.

B. Environment and Subjects

The study took place at a U.K. primary school towards the end of an academic year, in two matched age and ability classes, corresponding to the two conditions. A total of 59 children took part aged 7–8, 30 of whom were in the intervention condition of interest, and thus the primary focus of attention below (12 boys, 18 girls).

A robot and a 26" touchscreen (with supporting hardware) was placed in each of the classrooms permanently for a two week period. While in use during the school day there were no experimenters present; supervision was provided by class teacher, or teachers assistant. Interactions between a single child and the robot occured around the touchscreen (figure 1), which provides a mediator for the interaction – the context in which the child-robot interaction takes place.

C. Learning Task

The one-on-one interactions between the child and robot take place in the context of a broadly collaborative [8] sorting task, where the robot acts as a peer in attitude (e.g. informal, personalised, uses child's name) and knowledge (e.g. makes mistakes to the same extent as the child). This sorting task is centred on a large touchscreen [13], on which there are a set of images that need to be sorted into one of two categories (see figure 1 for an example). Each such set of images is labelled an "image library". In the present work, a total of 18 image libraries are used, with two subjects used (see table I): a familiar task for the children (maths times-tables), and an unfamiliar task (history – about the stone age). Each image library has an equal number of images for the two categories, with a total of 12 images for each maths image library and 14 images for the stone-age image libraries. The maths image libraries were organised such that there was a progressive increase in difficulty. This arrangement was verified with the class teachers prior to the study.

TABLE I
IMAGE LIBRARIES USED IN THE STUDY, SPLIT INTO FAMILIAR (MATHS TIMES TABLES) AND UNFAMILIAR (THE STONE AGE) TOPICS. NOTE THAT THE IMAGE LIBRARIES WERE INTERLEAVED DURING INTERACTIONS WITH THE ROBOT, AND THAT IN THE CASE OF THE FAMILIAR TOPIC, THE IMAGE LIBRARIES WERE ARRANGED IN INCREASING DIFFICULTY.

Maths (Familiar)		
Library	Contents	Task*
1	2x table	In/Out
3	10x table	In/Out
5	5x table	Odd/Even
7	2,10,5 div	Odd/Even
9	3x table	Odd/Even
10	4x table	In/Out
11	6x table	In/Out
12	3,4,6 div	Odd/Even
14	7x table	In/Out
15	8x table	In/Out
16	9x table	Odd/Even
17	11x & 12x	Odd/Even

Stone-Age (Unfamiliar)		
Library	Contents	Task*
2	SA lifestyle	Yes/No
4	SA animals	Yes/No
6	SA tools	Yes/No
8	SA art	Yes/No
13	SA mixed	Yes/No
18	SA mixed	Yes/No

* Task is a categorisation: the labels are the two categories shown on the screen

In this collaborative game setting, both the child and the robot have the same interaction affordances; i.e. they are both able to select an image, drag it, and deposit it in one of the category locations (see figure 1 for an example). There are no turn-based constraints, and overlapping actions on the touchscreen is possible – although in actual interaction, a turn-based structure does nevertheless appear to emerge from the interaction [14], indicating that in this context, the robot can be seen as a (potentially) social agent by the child.

Further supporting the notion that the robot was a peer, feedback to image categorisation moves on the touchscreen was provided visually on the screen itself (green tick or red cross): from the perspective of the child, the robot thus had the same feedback on performance that they had. The robot did however comment on the child's moves (e.g. "well done", or "maybe you'll do better on the next one"). No additional feedback information regarding individual images was provided: this is therefore a relatively sparse feedback regime. At the end of the image library (i.e. when all images had been sorted), and if the performance was below threshold, then the robot would make a brief comment (e.g. "oh dear, looks like the computer will make us do that one again") to indicate that a repeat would occur.

The main feature of the learning task with respect to the present paper is the possibility for repeating an image library if performance of the child is low. Since both the child and the robot are able to make categorisation moves on the touchscreen, we consider only the child's performance: i.e. only those moves made by the child on the touchscreen. Given that chance performance is 50% (two categories, equal number of members of each category), we consider acceptable performance to be at least 65% correct classifications (with a maximum number of three attempts). If the child's performance falls below this, then the library is reset once completed (i.e. a rearrangement of the same images on the touchscreen), up to a maximum of three times, after which the next image library would be shown. If an image library was completed successfully, then the next

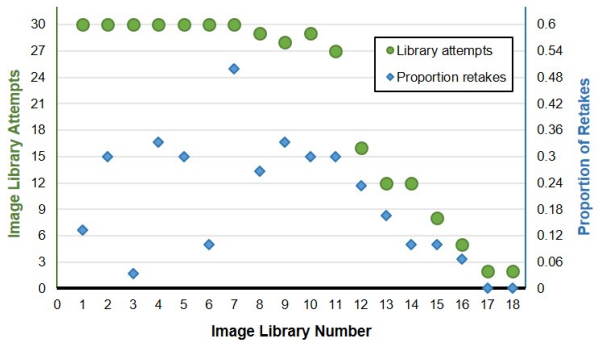


Fig. 2. Distribution of completions and proportion of repeats across image libraries for all children in the intervention condition.

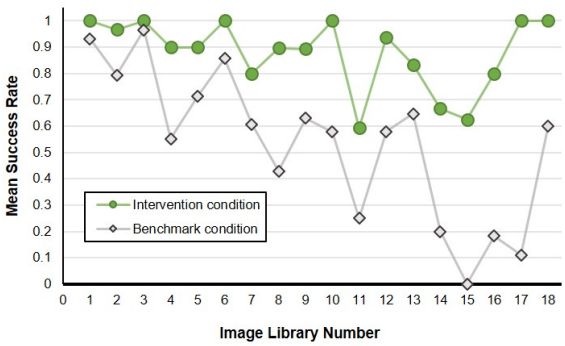


Fig. 3. Success rate for all image libraries in the intervention condition, and control condition. Success in a library is a child score of greater than 65% correct image classification.

image library (table I) would be automatically dispayed on the screen.

D. Procedure and Metrics

The hardware was set up in a corner of the respective classrooms at the start of the two-week experimental period, and remained in situ until the end. The system was started up each morning prior to the arrival of the children, and was shut down at the end of the school day after the children had left. No experimenters were present during the interactions of the children with the robot.

Over the course of the day, the teacher would nominate one child to interact with the robot at a time. This child would go over to the robot setup and interact while the rest of the class carried on with their normal activities. Each interaction would last five minutes (of interaction with the image libraries, not including introduction and closing procedures); over the course of the two week period, each child interacted with the robot on multiple occasions.

During each interaction, a range of information was collected. This included, for each child, the number of libraries completed, the child's score, and the number (and effect, in terms of score) of repeated image libraries. It is this data that is the primary subject of investigation below. In the wider study, a number of other metrics were recorded, including questionnaires, pre- and post-study knowledge tests, and video recordings – further details of these appear in [11].

III. Results

We reiterate at this point that the aim of this paper is to provide a descriptive analysis of data obtained that can be used as a basis for subsequent explorations, rather than as a hypothesis-led effort. Hence, while we make observations on a number of trends and relationships, we must leave further characterisation to future work. We further note that (unless otherwise stated), we focus on the results obtained in the intervention condition, i.e. the group of 30 children for whom there was the possibility of repeating image libraries.

A. Occurrence and Impact of Repeats

Not all of the libraries were completed by all 30 of the children over the course of the study (figure 2): after image library 11 (6x table), there is a sharp drop-off in completion rate. Considering the repeat rate for each library, it can be seen that there are a wide range of values. Where it may be expected that, for the maths libraries at least, increasing difficulty (seen in higher image library numbers) would result in a greater need for repeats, this is not evident from the data. A positive correlation is found here ($r = 0.914$, $n = 30$, $p < 0.001$), although this is likely to be due primarily to the drop-off in completion rate: those children likely to have progressed through more of the libraries may have been higher performing, hence requiring fewer repeats in the first place.

Repeating an image library does generally appear to confer an advantage in terms of score, when contrasted with a scenario in which no repetition is possible (figure 3). This provides an initial indication in support of the intuition that repetition of a task with a robot provides some advantage – however, due to the setup of the experiment, with a number of factors different between the conditions in addition to the possibility for repeats, this is not, on its own, conclusive.

In order to provide further insight, the impact of repeats per image library can be examined (figure 4). This shows that for most libraries where there are repeats, there is a score improvement from the first to the last attempt ($mean_{increase} = 0.218$, $n = 30$, 95% CI=[0.177,0.258]). The mean score change for the first image library seems to be an outlier here: it is likely to be due to uncertainty on the part of the four children as to what should be done; a shortfall quickly overcome on the second iteration. Indeed, each of these four individuals only had one repeat attempt.

B. Individual and Topic Differences

The overall difference in mean repeat rates between the maths libraries ($mean_{maths} = 0.424$, $n = 30$, 95% CI=[0.3,0.548]) and the stone age libraries ($mean_{SA} = 0.346$, $n = 30$, 95% CI=[0.217,0.475]) is small (with a large overlap in the 95% CIs). Examining the number of repeats per child across all image

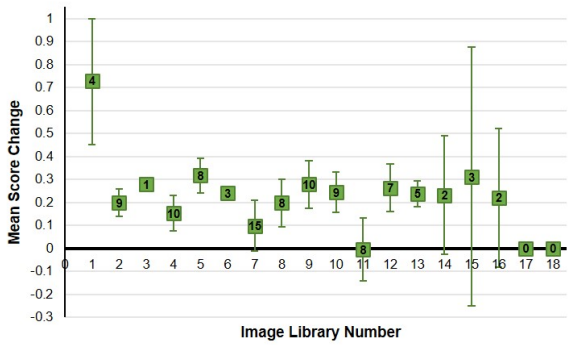


Fig. 4. Effect on score of repeat attempts, by image image library. Numbers in data points show number of repeats for that library across all children. Error bars show 95% CI.

libraries shows a high variability between children (figure 5). This seems to suggest that instead of looking at the group as a whole (i.e. is repetition generally a good strategy), it is necessary to consider the effects on individuals (i.e. under what circumstances and features of individuals does repetition confer a benefit to these individuals).

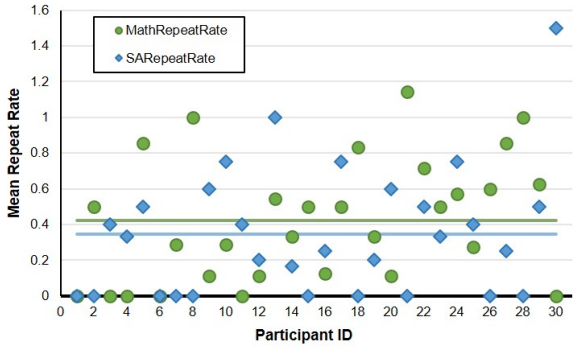


Fig. 5. Mean number of repeated attempts of image libraries per child, for the maths and stone age image libraries. Horizontal lines show mean for each image library subject.

This refocus on individual differences is further supported by considering the mean change in score achieved by each child (figure 6). While the difference in overall means is more pronounced between the image library subjects ($mean_{maths} = 0.132$, $n = 30$, 95% CI=[0.085,0.18]; $mean_{SA} = 0.076$, $n = 30$, 95% CI=[0.046,0.106]), a high degree of inter-subject variability is apparent[1]. Considering the relative performance increase for the two image library subjects, 18 children gained more from repeating maths image libraries, whereas only 10 individuals gained more from repeating the stone-age image libraries (two children did not repeat any image libraries).

[1]The mean values include values for those children who did not perform repeats. This is because repeat rate (or lack thereof) is a feature of the intersubject variability under examination; to exclude these instances would therefore be to skew the distribution under consideration.

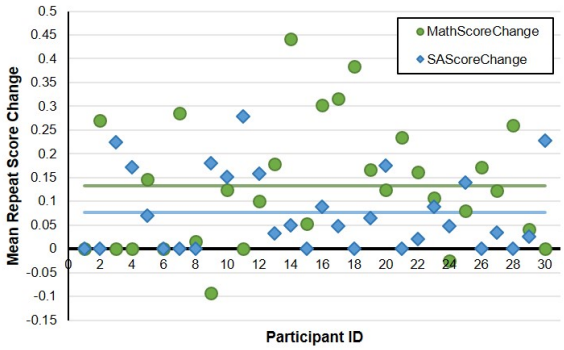


Fig. 6. Mean change in score after repeats per child, for the maths and stone age image libraries. Horizontal lines show mean for each image library subject.

TABLE II
CORRELATION MATRIX FOR MATHS TIMES TABLES (FAMILIAR) RESULTS. CELLS HIGHLIGHTED IN GREEN HAVE $p<0.05$, IN YELLOW IS $p<0.1$. N=30 FOR ALL CORRELATIONS. *Perf*: OVERALL CHILD CLASSIFICATION PERFORMANCE. *Rep rate*: MEAN REPETITION RATE. *Tot reps*: TOTAL NUMBER OF REPEAT ATTEMPTS. Δ*Score*: CHANGE IN SCORE, PRE- TO POST-REPEAT. *N_libs*: TOT NUMBER OF IMAGE LIBRARIES COMPLETED.

	Gender	Perf	Rep rate	Tot reps	ΔScore	N_libs
Gender	1					
Performance	-0.075	1				
Rep rate	0.124	-0.760	1			
Tot reps	0.100	-0.770	0.960	1		
ΔScore	0.278	-0.236	0.450	0.402	1	
N_libs	0.192	0.321	-0.245	-0.239	-0.125	1

TABLE III
CORRELATION MATRIX FOR STONE-AGE (UNFAMILIAR) RESULTS. CELLS HIGHLIGHTED IN GREEN HAVE $p<0.05$, IN YELLOW IS $p<0.1$. N=30 FOR ALL CORRELATIONS. LABELS AS FOR TABLE II

	Gender	Perf	Rep rate	Tot reps	ΔScore	N_libs
Gender	1					
Overall Perf	-0.075	1				
Rep rate	0.255	-0.112	1			
Tot reps	0.072	-0.182	0.931	1		
ΔScore	0.138	0.289	0.583	0.584	1	
N_libs	0.183	0.313	0.226	0.133	0.377	1

C. Indications from Correlations

In order to explore what individual influences there are on the effect of repeating image libraries on performance within the context of this study, we explore correlations between the various metrics recorded during the study. This form of analysis naturally does not provide proof of causality, but it can provide indications of trends, and relationships that could be explored further. First we break this down by image library subject (tables II and III), before considering the relationship between the two (table IV).

As would be expected, there is a strong (and significant) association between repeat rates and total number of repeats. Similarly, and in support of figure 4, there is a strong and statistically significant association between repeat rate (and total number of repeats) and score change, for both image library subjects: i.e. the greater the number of repeats, the greater the change in score.

However, one clear difference between the correlations for

TABLE IV

CORRELATION MATRIX COMPARING MATHS AND STONE-AGE RESULTS. CELLS HIGHLIGHTED IN GREEN HAVE $p<0.05$, IN YELLOW IS $p<0.1$. N=30 FOR ALL CORRELATIONS. *Math/SA-Libs*: TOTAL NUMBER OF IMAGE LIBRARIES ATTEMPTED OF RESPECTIVE SUBJECTS. *Maths/SA-Re*: NUMBER OF REATTEMPTS FOR EACH RESPECTIVE SUBJECT.

	Perf	SA-Libs	SA-Re	ΔSA	SA-Success
Perf	1	0.313	-0.182	0.289	0.214
Math-Libs	0.321	0.923	0.134	0.333	-0.075
Maths-Re	-0.770	-0.297	-0.261	-0.663	-0.120
ΔMath	-0.236	-0.083	-0.295	-0.445	-0.160
Maths-Success	0.644	0.320	0.067	0.323	-0.031

the two image library subjects is in the relationship between the repetition rate (and total repeats) and the overall image library performance (mean per child over the whole study period): for the maths times tables image libraries this is a strong negative correlation (significant), whereas this is only weak (non-significant) for the stone-age image libraries.

Considering the relationships between maths and stone-age image library-related behaviour provides some further insight into individual differences (table IV). Firstly, as would be expected, the number of maths and stone-age image libraries completed is strongly positively correlated. Secondly, there is a strong positive correlation between the overall performance and maths image library success rate, but not for the stone-age image library success; this is despite there being an overall higher success rate for the stone-age ($mean_{SA} = 0.934$, $n = 30$, 95% CI=[0.887,0.982]) than maths image libraries ($mean_{maths} = 0.862$, $n = 30$, 95% CI=[0.816,0.908]). Thirdly, there is a moderate negative (significant) correlation between change in stone-age performance after repeats and both number of math repeats and change in math score. Furthermore there is a strong negative correlation between overall performance and number of math image library repeats (the more repeats needed, the lower the overall score, and vice versa). The presence of only a few significant results here make patterns and trends difficult to extract, but in general the results seem to suggest that performance and change in performance is inversely associated for the two image library subjects.

One final aspect to note regarding the separate image library subject correlations is that the gender of the child does not appear to be strongly (or significantly) associated with any of the other variables. For this reason, the effect of gender is not considered further for the present paper (although there may be related phenomena worth further investigation).

IV. DISCUSSION

A central facet of the experimental setup and task context used is that it is a fundamentally collaborative task between a child and a robot (figure 1). Note that despite collaboration not being enforced (i.e. rather than having an explicit turn-taking structure, it is possible for the child to complete the task on their own if he/she ignores the robot), collaborative behaviours are indeed typically observed [14]. It is in this interactive context that the results obtained should be considered. With the robot taking on the role of a peer (see section II-C), the extent of feedback provided to the child is relatively sparse (owing to the desire for the robot to have the same level of apparent knowledge as the child). Nevertheless, this feedback serves to highlight to the child where image libraries are to be repeated: any subsequent change in performance (such as the mean increase observed in the present study) may thus be mediated by this interactive context. The collaborative nature of the task may also provide additional motivation for performance improvement (beyond the desire to move on to another image library) [15], although this effect requires further empirical investigation.

The results have shown that at the group level, there is some apparent benefit for repeating a categorisation task if initial performance is low (figure 4), and that this benefit is greater for the familiar subject than the unfamiliar one (figure 6). Familiarity may in this case not be the only distinguishing characteristic between the two types of image library, with other aspects such as the level of abstraction or topic-related enjoyment that may be important: this requires further investigation, although we note that levels of self-reported enjoyment remain high [11]. However, it is also clear that (as may perhaps be expected) there is a high degree of variability between subjects. Examining these more closely indicates that repeats for maths is positively correlated with score change, but inversely correlated with overall performance – a relationship not present for the stone-age subject.

One feature of the results is that both the overall repeat rate and the overall score change as a result of repeats is higher for the familiar subject (maths) than for the unfamiliar subject (stone-age). We suggest that this may be related to the sparsity of the feedback: recall that correct/incorrect feedback is only provided on the touchscreen itself in response to a categorisation. In a familiar task, the children would already know the features of the problem (what is involved in multiplication for example), and so even sparse feedback is confirmatory. Conversely, this may not be true for a novel problem, in which case only sparse feedback may not be as helpful. This seems to be supported by the correlation results, where there was a strong negative correlation between repeat rates and overall performance for the familiar subject, but not for the unfamiliar subject. This leads to a hypothesis for future study that for unfamiliar tasks (to the children), richer feedback is required than for familiar tasks.

Note however that the relationship between the group data and the correlations remains ambiguous in some respects. For example, the negative correlation between change in math performance and change in stone-age performance (table IV) requires further investigation. It is likely that the wider context for the individual needs to be taken into account, as in the discussion of sparse feedback and role of interactivity above. One possibility not explored in the present study is the role of attention: repetition could be a means of re-orienting attention back to the task after a lapse of concentration or misunderstanding (cf. the outlier mean score change in the first image library, figure 4). More generally, the question is – what characteristics of the individual (or the circumstances) predispose them (or not) to gain more from repetition?

This widening of scope seems necessary when performing a more fine-grained analysis. Returning to the notion of repeating collaborative tasks based on initial performance, we have seen that while in general there could be some benefit, it is necessary to consider this from the perspective of the children's individual differences and of the task engaged in (familiar and unfamiliar in this case). While the present study has only provided a descriptive analysis of the data obtained, it provides a number of pointers to phenomena that should be further researched.

ACKNOWLEDGMENT

This work was supported by the EU FP7 project DREAM (grant number 611391, http://dream2020.eu), and the H2020 project L2TOR (grant number 688014, http://www.l2tor.eu). The authors would like to thank Salisbury Road Primary school (Plymouth, U.K.) for their participation in the study.

REFERENCES

[1] F. Tanaka and S. Matsuzoe, "Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 78–95, 2012.

[2] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayahuitl, B. Kiefer, S. Racioppa, I. Kruiff-Korbayova, G. Athanasopoulos, V. Enescu, R. Looije, M. Neerincx, Y. Demiris, R. Ros-Espinoza, A. Beck, L. Canamero, A. Hiolle, M. Lewis, I. Baroni, M. Nalin, P. Cosi, G. Paci, F. Tesser, G. Sommavilla, and R. Humbert, "Multimodal Child-Robot Interaction : Building Social Bonds," *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 33–53, 2012.

[3] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Modelling Empathic Behaviour in a Robotic Game Companion for Children : an Ethnographic Study in Real-World Settings," in *HRI'12*. Boston, MA, U.S.A.: ACM Press, 2012, pp. 367–374.

[4] R. Ros, I. Baroni, and Y. Demiris, "Adaptive humanrobot interaction in sensorimotor task instruction: From human to robot dance tutors," *Robotics and Autonomous Systems*, vol. 62, no. 6, pp. 707 – 720, 2014.

[5] S. Lemaignan, A. Jacq, F. Garcia, D. Hood, A. Paiva, and P. Dillenbourg, "Learning by teaching a robot: The case of handwriting," *IEEE Robotics Automation Magazine*, vol. PP, no. 99, pp. 1–1, 2016.

[6] B. Skinner, "The science of learning and the art of teaching," *Harvard Educational Review*, vol. 24, pp. 86–97, 1954.

[7] A. S. Palincsar, "Social constructivist perspectives on teaching and learning," *Annual Reviews of Psychology*, vol. 49, pp. 345–375, 1998.

[8] P. Dillenbourg, "What do you mean by collaborative learning?" in *Collaborative Learning: Cognitive and Computational Approaches*, P. Dillenbourg, Ed. Elsevier, 1999, pp. 1–15.

[9] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial," *Human-Computer Interaction*, vol. 19, no. 1, pp. 61–84, jun 2004.

[10] J. Kennedy, P. Baxter, and T. Belpaeme, "Comparing Robot Embodiments in a Guided Discovery Learning Interaction with Children," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 293–308, 2015.

[11] P. Baxter, E. Ashurst, R. Read, J. Kennedy, and T. Belpaeme, "Robot education peers in a situated primary school study: Personalisation promotes child learning," *PLOSONE*, in review.

[12] P. Baxter, E. Ashurst, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme, "The Wider Supportive Role of Social Robots in the Classroom for Teachers," in *1st Int. Workshop on Educational Robotics at the Int. Conf. Social Robotics*, Paris, France, 2015.

[13] P. Baxter, R. Wood, and T. Belpaeme, "A Touchscreen-Based Sandtray' to Facilitate, Mediate and Contextualise Human-Robot Social Interaction," in *7th ACM/IEEE International Conference on Human-Robot Interaction*. Boston, MA, U.S.A.: IEEE Press, 2012, pp. 105–106.

[14] P. Baxter, R. Wood, I. Baroni, J. Kennedy, M. Nalin, and T. Belpaeme, "Emergence of Turn-taking in Unstructured Child-Robot Social Interactions," in *HRI'13*, no. 1. Tokyo, Japan: ACM Press, 2013, pp. 77–78.

[15] A. Coninx, P. Baxter, E. Oleari, S. Bellini, B. Bierman, O. B. Henkemans, L. Canamero, P. Cosi, V. Enescu, R. R. Espinoza, A. Hiolle, R. Humbert, B. Kiefer, I. Kruijff-korbayova, R. Looije, M. Mosconi, M. Neerincx, G. Paci, G. Patsis, C. Pozzi, F. Sacchitelli, H. Sahli, A. Sanna, G. Sommavilla, F. Tesser, Y. Demiris, and T. Belpaeme, "Towards Long-Term Social Child-Robot Interaction: Using Multi-Activity Switching to Engage Young Users," *Journal of Human-Robot Interaction*, vol. 5, no. 1, pp. 32–67, 2016.

The Cautious Attitude of Teachers Towards Social Robots in Schools

James Kennedy, Séverin Lemaignan, Tony Belpaeme
Centre for Robotics and Neural Systems
Plymouth University, U.K.
Email: {firstname.surname}@plymouth.ac.uk

***Abstract*—Social robots are increasingly being applied in educational environments such as schools. It is important to understand the views of the general public as social acceptance will likely play a role in the adoption of such technology. Other literature suggests that teacher attitudes are a strong predictor of technology use in classrooms, so willingness to engage with social robots will influence application in practice. In this paper we present the results of a rigorously-framed survey used to gather the views of both the general public and education professionals towards the use of robots in schools. Overall, we find that the attitude towards social robots in schools is cautious, but potentially accepting. We discuss the reported set of perceived obstacles for the broader adoption of robots in the classroom in this context. Interestingly, concerns about appropriate social skills for the robots dominate over practical and ethical concerns, suggesting that this should remain a focus for child-robot interaction research.**

I. Introduction

Research involving social robots in educational settings is becoming increasingly prevalent, particularly with children [1], [2]. Indeed, researchers in established fields applied to the educational domain, but using different technologies, have started to call for a switch to developing and evaluating social robots [3]. Work conducted within the field of Human-Robot Interaction (HRI) is taking place over longer-term time-scales as well, inspired by early success stories such as [4], and striving for increasingly sustained real-world application.

It has been shown that robots can be used to successfully teach children, and also offer unique learning experiences. For example, children can teach a less-able peer (in the form of a robot), which may not otherwise have been possible [5], [6]. However, they can also have an impact on the classroom, both in terms of the child behaviour and teacher behaviour [7] (which is also related to the broader concept of technology-mediated *classroom orchestration* [8]).

As this field of research pushes forwards, and if we seek further real-world or mass-market implementation in schools, it is important to understand attitudes towards the technology. For successful adoption of such technologies, it is necessary for both teachers and the general public to be willing participants in increased uptake. Recent findings from the Eurobarometer report [9] have suggested that whilst there is generally a positive view towards robots in Europe, there is a sizeable contingent (34%) that would see robots banned from use in education. However, the survey administered in this report does not provide a context for many of the questions.

In this paper we seek to explore whether, when provided a minimal context, the attitudes of the general public are in fact more positive. We explore the impact of this context on the responses by manipulating an 'imagined' picture of how a classroom with a robot might look (by including a human teacher or not). Using the same survey design we also seek to establish views of teachers (for whom there will be a greater direct impact) regarding the use of social robots in education. Furthermore, the views of teachers about obstacles to the use of robots are considered for insight into possible child-robot interaction research directions.

II. Related Work

Research has suggested that there are barriers to adoption and use of technology by teachers. These can be first-order (extrinsic) barriers, or second-order (personal) barriers. While the extrinsic barriers cannot be discounted, it has been found that positive beliefs of teachers about the effectiveness for learning (i.e., personal factors) are a significant predictor of actual technology use [10]. For this reason, it is important to understand (and possibly influence) how teachers feel towards social robots if we intend to see them widely adopted. Teacher views may also highlight research questions that need to be addressed to demonstrate the efficacy and suitability of using robots in schools.

Previous pan-European work [11] found that views of teachers are generally positive, but that there are concerns over fairness to access, the robustness of the technology, and potential disruption to classrooms. Some of these same concerns were observed prior to an experiment in the USA, but after the experiment had been completed, views had changed [12]. Teachers expected the robot to be disruptive to the classroom, but found that it was not, although this is partially mitigated as headphones were used so that the possibility of audible disruption would be minimised. A large-scale survey conducted in South Korea [13] found that teachers were generally positive about the use of robots in education, but they were more negative than other stakeholders. Ethical tensions have also been identified pertaining to issues of privacy, robot role, socio-emotional effects on children and responsibility [14].

When exposed to a highly scripted interaction with a robot, teachers showed fairly positive reactions [15], however it was concluded that the interaction here was not related to the educational quality that the robot could offer, and this is

Fig. 1. 'Imagined' classroom with the human teacher present. This is used on the survey in the 'teacher' (TE) condition.

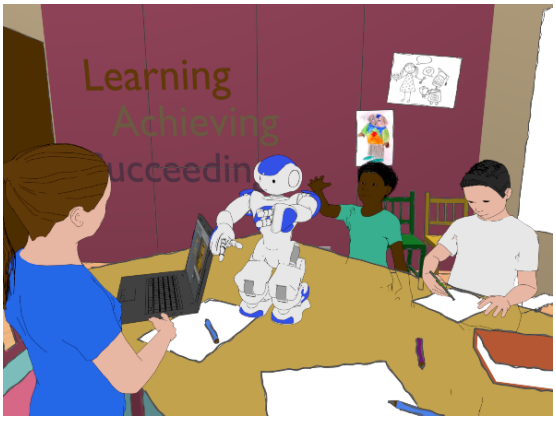


Fig. 2. 'Imagined' classroom without the human teacher present. This is used on the survey in the 'no teacher' (NT) condition.

where the focus should be. Incorporating the views of teachers in educational technology design has been highlighted as a particularly important aspect of creating a partnership that allows teachers to identify the benefits and shortcomings of technology when related to the curriculum [16]. This motivated us to consider how we might gather the opinions of both the general public and education professionals, with the aim of using the findings to direct future research.

Due to the technological nature of robots, it is anticipated that they will be seen as a tool for STEM education, rather than for the teaching of humanities. This is reflected in the research being conducted with robots in education: they are commonly applied in STEM education, with promising outcomes [17], although research is also prominent in language contexts [1], [4]–[6]. However, there are comparatively few robots being used to teach art or religious education, for instance (a reference to work in either of these domains could not be identified at the time of writing). These pre-conceptions will be explored as they could produce further barriers to adoption of the technology in certain areas (or indeed may highlight areas that should not even be attempted to be addressed with robots).

III. Hypotheses

From the related work outlined in the previous section and our prior experience, the following hypotheses were devised for this study:

- H1 *Context matters:* providing a minimal context will lead to more positive attitudes towards robots in education than the Eurobarometer [9] suggests.
- H2 *Robots for STEM:* robots will be seen as an educational tool for delivering science, technology, engineering and maths (STEM) content, but not for broader use in the arts or humanities.

Additionally, we seek to address the following exploratory question to build on prior research [11], [12], [14]: Q1 'what are some potential obstacles perceived by educators to the adoption of robots in the classroom and what can be done by researchers regarding these?'.

IV. Methodology

A. Survey Design

In order to gather the opinions required to address the hypotheses, we devised a survey to elicit the attitudes of people towards the use of social robots in education. Part of this survey was based on the questions asked in the Eurobarometer survey [9], whilst other questions were devised by the authors to specifically focus on areas of interest relating to the hypotheses and applications of robots in education. The full survey is not included here due to space restrictions, but can be viewed online: https://github.com/james-kennedy/r4lworkshop-survey.

Two versions of the survey were created: (1) with a picture with a teacher present (TE), and (2) without a teacher present (NT; Fig's. 1 and 2). This was done as a methodological check to explore whether the image provided to participants would shape their attitudes towards robots in schools. In both cases, the accompanying text was kept the same: a broad description of social robots and of their abilities in relation to learning ('the children can talk to the robots and learn from them', 'the robot can learn children's names and preferences', 'it can personalise learning experiences').

B. Participants

Two pools of participants were recruited to address the hypotheses: (1) education professionals from schools in the U.K., and (2) members of the general public. The members of the general public completed an online questionnaire via a crowdsourcing platform (http://www.crowdflower.com). The online responses were limited to the top 2 levels (indicating 'extremely high' previous response quality) of 'contributor' as judged by the crowdsourcing platform. Respondents were restricted to the U.K. (to match the education professionals country). All participants consented to having their responses used for research purposes. The general public were compensated with an amount commensurate with the national living wage at the time of execution; the educators received no compensation.

General public (GP): 100 responses were collected; 50 with each picture. The responses were manually checked and it was found that some responses were from the same users with multiple accounts (6 instances), whilst others were in fact from those working in education (7 instances). These responses were therefore removed, leaving a total of 87 responses (41 TE/46 NT). The average age of this sample was 35.3 years (*SD*=11.4), 29F/58M. Further demographic details (such as number of children and education level) were collected and will be explored as factors in the analysis in Sec. V.

Education professionals (EP): 35 responses were collected (19 TE/16 NT). The average age was 37.6 years (*SD*=11.5), with 2 not providing their age. The sample has a strong female bias (31F/4M), which reflects the gender balance in the U.K. for primary school employees. We focus on primary schools as this is the age commonly used in HRI research in education settings. The sample came from two schools; one in a rural location (18 responses), and one in a city (17 responses). Both class teachers and teaching assistants were included.

V. Results

Preliminary analysis was conducted to verify the reliability of the data. Cronbach's Alpha was calculated for an 8 item sub-scale of the survey that related to the acceptance of robots in education (questions 4 to 10 and 14). This was performed on 98 of the 122 total responses (due to non-responses or 'unsure' responses), resulting in $\alpha = .862$. This value indicates that the internal consistency of responses is high, so the data is likely to be reliable.

To test the stimulus manipulation, a comparison within each of the groups (EP and GP) was performed between those who had seen the survey with the teacher in the picture and those without the teacher. For this, Mann-Whitney *U* tests were conducted for the questions relating to acceptance of robots in education (the same ones as for Cronbach's Alpha: questions 4 to 10 and 14). No significant differences were found for any of the questions for the GP sample (*U* values varied from 666.5 to 904.0 and *p* values varied between .161 and .731). Nor were significant differences found for the EP sample ($49.0 < U < 140.5$; $.142 < p < .712$). This provides a strong indication that the change in picture stimulus did not cause significant differences in responses. Due to this, for the remaining analysis, no distinction will be made between the two conditions with (TE) and without (NT) teacher visible in the stimulus.

A. Interest in Technology and Positivity Towards Robots

When seeking to address Hypothesis 1, we identified a bias towards having a favourable view of technology in the data collected from the online survey. The first question of the survey asks how interested the participant is in science and technology (*very*, *moderately*, or *not at all*). For the EP, the split falls roughly in line with that of the Eurobarometer [9], but our general public view is clearly more interested (Table I). This is reflected in a comparison between the general public (Mdn=3) and educator (Mdn=2) responses using a Mann-Whitney test: $U = 1029.5, p = .002, r = .29$. This also carries through to how positive a view they hold about social robots (question 2; 5 point Likert from *very negative* to *very positive*). A Mann-Whitney test indicated that the general public held a more positive view of social robots (Mdn=4) than educators (Mdn=3), $U = 820, p = .001, r = .32$.

TABLE I
Interest in science and technology as reported by survey respondents (and the Eurobarometer [9]).

Group	Very interested (%)	Moderately interested (%)	Not at all interested (%)
General public	61	37	2
Educators	31	57	12
Eurobarometer	25	47	28

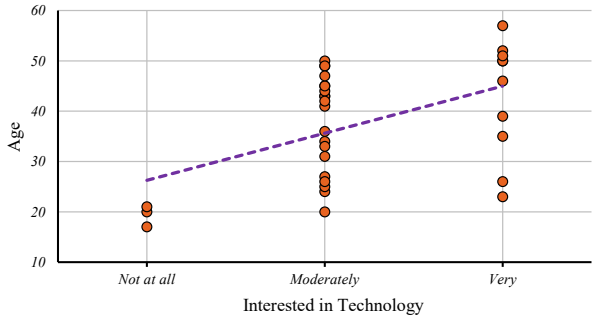


Fig. 3. A significant correlation is observed between educator age and interest in technology, with younger educators reporting to have less interest in technology.

These responses were correlated with the questions regarding views about the use of robots being used in education. It was found that a positive correlation exists between how positive a view someone has about social robots (question 2) and the role that a robot should play in child education for both educators ($r_s(25) = .561$, $p = .002$) and the general public ($r_s(84) = .390$, $p < .001$). These fundamental differences cause problems in comparing between educators and the general public, and the general public and the Eurobarometer findings. If it were reflective of differences between the general public and educators, then this would be an acceptable factor, but we hypothesise that it is instead because of a pro-technology bias caused by the online method used to gather general public responses. As such, a direct comparison would not be appropriate for exploring Hypothesis 1, nor can the EP and GP samples be considered homogeneously.

There is an observed positive correlation between age and interest in technology for educators ($r_s(31) = .492$, $p = .004$; Fig. 3), but not for the crowdsourced responses ($r_s(85) = -.093$, $p = .393$). This is probably due to the self-selecting nature of the crowdsourced participants, but is an interesting finding for the educators – this will be returned to in the discussion (Sec. VI).

Due to the differences between our crowdsourced sample and the Eurobarometer sample, a direct comparison that was intended to be explored as part of Hypothesis 1 (that providing

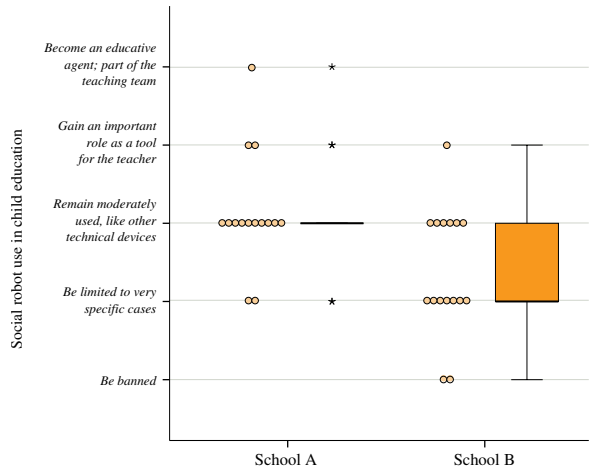


Fig. 4. Opinions from educators about how robots should ideally be used in child education split by school. This was a forced choice survey item, with an implicit scale from 1 to 5: 'be banned', 'be limited to very specific cases', 'remain moderately used, like other technical devices', 'gain an important role as a tool for the teacher', 'become an educative agent; part of the teaching team' (and an 'I don't know' option, not shown). * indicates outliers.

a context as we do in our survey will lead to more positive responses) would not be sound. However, it should be noted that the Eurobarometer reporting of 34% wanting robots to be banned in education was not reflected in our results, where only 2 respondents (both from the educator sample) want robots to be banned from use in education (Fig. 4).

B. Cultures Within Schools

To further explore the views of the education professionals, we compared the responses from the different schools. We find that despite there being no significant differences in interest in technology (School A: Mdn=2, School B: Mdn=2; Mann-Whitney $U = 123, p = .263, r = .19$), there are differences in attitudes towards the use of social robots in education. Question 14 on the survey (see Fig. 4) is particularly indicative of an overall view, asking how social robots should ideally be used in child education. These answers were converted to an ordinal scale, with *be banned* receiving the lowest score, and *become an educative agent; part of the teaching team* the highest. A Mann-Whitney U test found that a significant difference exists between School B (Mdn=2) and School A (Mdn=3), $U = 62, p = .012, r = .45$ (Fig. 4).

No significant demographic differences could be found between the two schools to explain the difference in attitudes, although their locations could be a factor. School A, which appears to be more open to the use of social robots in education is situated in a rural village (population approx. 7,000), whereas School B is within a reasonably large U.K. city (population approx. 250,000). We would hypothesise two possible explanations: (1) differing micro-cultures between large cities and small villages lead to different concerns for children's well-being, or (2) differing ethos between schools regarding their attitude in general towards teaching science and technology. The former will be discussed further in Sec. V-D, but the latter would require further investigation to analyse the 'culture' within the schools.

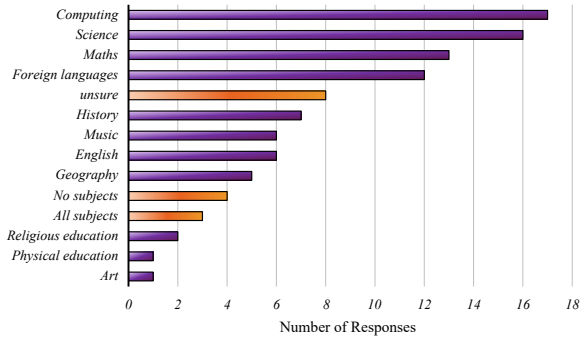


Fig. 5. Opinions from education professionals about the subjects in which they think social robots could be used to aid learning (forced choice survey item; multiple responses can be selected, leading to 101 total responses).

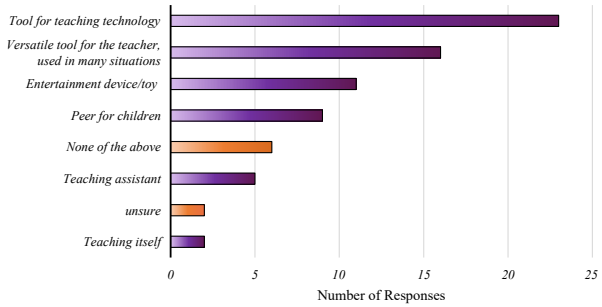


Fig. 6. Opinions from education professionals about how robots could be used in a school classroom (forced choice survey item; multiple responses can be selected, leading to 74 total responses).

C. Robots as a STEM Tool

Two questions on the survey were used to address how people perceived the uses of robots in terms of the content it could deliver, and in which role (Hypothesis 2). It was hypothesised that robots would be seen as a tool for delivering STEM education, and indeed this was supported through the data. Twenty of the 35 educators thought that the robot could be used to aid learning in computing (which covers programming, I.T., digital security, etc.), followed by science (19) and maths (16), with humanities such as art (4) and religious education (5) receiving very few responses (Fig. 5).

The survey question 11 asked about the envisioned role of social robots in the classroom, with several options ranging from an 'entertainment device', a 'tool', a 'peer for children', and a 'teacher itself' (see Fig. 6 for all options). In line with the results presented in Fig. 4 and in the previous paragraph, the education professionals mainly see robots as tools (Fig. 6) – again providing support for Hypothesis 2. In more than 30% of the cases, the EP also view the robot as a toy, which may reflect misconceptions or a lack of clarity about robots in a learning environment. We comment further on this point in the discussion.

TABLE II
PERCEIVED OBSTACLES TO ADOPTION, AS MENTIONED IN FREE TEXT ANSWERS TO QUESTION 15. PARTICIPANTS COULD MENTION SEVERAL ITEMS. THE PERCENTAGE OF RESPONDENTS MENTIONING THE ITEM IS PROVIDED FOR BOTH EDUCATION PROFESSIONALS (EP) AND THE GENERAL PUBLIC (GP) WITHIN EACH GROUP.

Obstacle	**#EP**	**% of cases**	**#GP**	**% of cases**
Source of distraction	10	34.5%	10	16.1%
Lack of social skills	9	31.0%	9	14.5%
Practical issues	7	24.1%	17	27.4%
of which, cost	*1*	*3.4%*	*12*	*19.4%*
Risk of isolation	6	20.7%	1	1.6%
Workload/orchestration load	5	17.2%	6	9.7%
Public perception	2	6.9%	10	16.1%
Ethical concerns	2	6.9%	1	1.6%
Safety	1	3.4%	2	3.2%
Technical limitations	1	3.4%	7	11.3%
Educational efficacy	0	0.0%	9	14.5%
Societal impact	0	0.0%	8	12.9%

D. Perceived Obstacles to Adoption

To explore Question 1 (Sec. III), a question was used to ask 'what would you see as the main obstacles for having robots in a classroom?'. This question had a free text answer so that responses were not constrained; an answer was not forced for this question. The responses from the educators provided many insights into the use of social robots in schools, often revealing deeper concerns that were hard to capture through other questions. Of the 35 EP respondents, 29 provided an answer for this question, and of the GP respondents, 62 provided an answer. We group these responses in a series of categories (formed by considering all responses), which are shown in Table II.

The most cited obstacle to adoption for EP is the robot being a potential source of distraction for the children – something that falls in line with prior research [11], [12]. However, this rather broad category could actually reflect the fact that teachers do not have a clear idea of what the robots could be used for (the context provided for the survey was minimal, so a precise role for the robot was not specified). In contrast, the most cited obstacle perceived by the GP sample were practical issues, and in particular, the cost of the robot. Cost was not mentioned in the survey at any stage, so this indicates that there is a pre-conception that these robotic devices would be expensive (or at least more expensive than schools can afford).

The perceived lack of social skills (simplistic interactions, lack of empathy, lack of flexibility) of robots gives a complementary picture of the current perception of robots by the education professionals: they are primarily seen as a scripted, reactive machine. This issue was somewhat surprising as it had not commonly been raised as an issue in prior work. More expectedly, a range of practical issues (cost, maintenance, space requirements) are mentioned, but usually along with other factors. Contrary to the perception by the general public, they do not appear to be the teachers' main concern at this stage.

Another factor that had not been hypothesised was the mention by several teachers of an increased risk of child isolation (for example, one comment read: 'I consider that many of our children are already isolated and this could isolate and potentially marginalise them further'). This would support the pushing forward of social approaches to child-robot interaction, like robot-mediated collaborative learning (i.e., using technology to further encourage interactions between child peers).

Some concerns were also raised in relation to the increased workload or classroom orchestration load brought by the robots for the teachers. These issues have been studied in the context of computer-supported learning (for instance [18]), but are yet to be fully considered in the field of 'robot-supported' learning.

Finally, surprisingly few ethical and safety-related concerns were raised. Such concerns do not appear to be prevalent amongst the EP respondents.

E. Demographic Factors

Other demographic factors in the education professionals sample (age, gender, number of children, education level) do not appear to have an impact on opinions about how social robots should be used in child education. Linear ordinal regression does not reveal a statistically significant factor when considering participant age, gender, number of children, or education level (Nagelkerke pseudo $R^2 = .146$, so the demographic factors only account for around 15% of the variance in how participants believe social robots should be used in child education). A model with a high goodness-of-fit could not be found when performing the same regression on the data from the general public (possibly due to the sample bias towards high interest in technology overpowering the other factors).

VI. DISCUSSION

A bias towards a positive view of science and technology was introduced through the means of collecting responses from the general public - via an online crowdsourcing service. This prevented us from directly addressing Hypothesis 1 through a comparison to the Eurobarometer survey data. However, we do see that there is a general openness to using social robots in education, although education professionals may approach this with a degree of caution (Fig. 4, Sec. V-D). There is also a strong pre-conception from educators that social robots would be suitable for teaching STEM subjects, adopting the role of a tool, rather than as an educative agent (Hypothesis 2, Sec. V-C, Fig. 5). These findings were observed regardless of whether respondents had been presented with a picture including a teacher, or not including a teacher in the introductory context for the survey (Sec. V).

Some perceptions based on pre-conceptions may well change with greater exposure to social robots that can do more than be used as a tool for STEM subjects (for example, as recently shown with handwriting learning [5]). However, a general lack of interest in science and technology (particularly from younger educators – Sec. V-A) could produce greater, and cyclical barriers to use. It has been shown that there are links between teacher interest and confidence in teaching subjects [19], as well as reciprocal effects between teachers

and child in engagement in learning [20]. It follows that if teachers are less interested in teaching technology, students will be reciprocally less interested, they will learn less [21], and be less likely to continue study of that subject [22]. This presents a concerning cycle wherein those students who eventually become teachers are also likely to lack interest in teaching those same subjects. The lack of interest of younger teachers for technology also comes as a surprise as one would typically expect younger teachers to be more engaged with computer-related technologies.

This is potentially where the broader aspects of using a social robot could be beneficial in breaking down some barriers to use. The robot is a technological device, but could be used to teach a variety of subjects with an element of sociality. The use of the robot could stimulate interest in technology, and the social aspects of robot behaviour could be used to create reciprocal interest in those subjects (as has been attempted for some aspects of behaviour [23]). This calls for a greater exposure of teachers to our robotic systems, so that they better comprehend the capabilities, current limited performance, and possible future applications of social robots in education.

Successfully addressing the concerns highlighted by educators in Sec. V-D (in relation to Question 1, Sec. III) would provide an essential first step towards this goal. Some of the concerns may arguably be alleviated once the teachers (and the children) familiarise themselves with the robots (the robot being a source of distraction is likely to resolve quickly after novelty goes away) or once the penetration of robots in classrooms increases to a point where dedicated companies could regularly take over training and maintenance issues. However, other issues, like the richness of the interaction, the adaptability of the robots to rapidly (or, on the contrary, slowly) change in response to child behaviours, or the suitability of social robots to develop children's peer-group sociality, present more fundamental questions. We believe that these behavioural considerations must remain central to the research agenda of child-robot interaction.

VII. Conclusion

Overall, we find that the attitude towards social robots in schools is cautious, but potentially accepting (in line with previous findings [13]). The perceived obstacles to adoption of robots in classrooms which the education professionals highlight raised some surprising considerations, such as potential isolation of students which would warrant further long-term study. For the educators, concerns about appropriate social skills for the robots dominate over practical and ethical concerns, suggesting that this should remain a focus for child-robot interaction research.

Acknowledgements

This work has been partially supported by the EU H2020 L2TOR project (grant 688014), the EU FP7 DREAM project (grant 611391), and the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227).

References

[1] J. Kennedy *et al.*, "Social Robot Tutoring for Child Second Language Learning," in *Proc. of the 11th ACM/IEEE Int. Conf. on HRI*. IEEE Press, 2016, pp. 67–74.

[2] A. Ramachandran *et al.*, "Shaping productive help-seeking behavior during robot-child tutoring interactions," in *Proc. of the 11th ACM/IEEE Int. Conf. on HRI*. IEEE Press, 2016, pp. 247–254.

[3] M. J. Timms, "Letting artificial intelligence in education out of the box: Educational cobots and smart classrooms," *International Journal of Artificial Intelligence in Education*, pp. 1–12, 2016.

[4] T. Kanda *et al.*, "Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial," *Human-Computer Interaction*, vol. 19, no. 1, pp. 61–84, 2004.

[5] D. Hood *et al.*, "When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting," in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI*. ACM, 2015, pp. 83–90.

[6] F. Tanaka and S. Matsuzoe, "Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 78–95, 2012.

[7] P. Baxter *et al.*, "The wider supportive role of social robots in the classroom for teachers," in *Proc. of the 1st Int. Workshop on Educational Robotics, at ICSR'15*, 2015.

[8] P. Dillenbourg and P. Jermann, *New Science of Learning: Cognition, Computers and Collaboration in Education*. Springer New York, 2010, ch. Technology for Classroom Orchestration, pp. 525–552.

[9] "Special Eurobarometer 382: Public Attitudes Towards Robots," European Commission, Tech. Rep., 2012. [Online]. Available: http://ec.europa.eu/public_opinion/archives/ebs/ebs_382_en.pdf

[10] C. K. Blackwell *et al.*, "Adoption and use of technology in early education: The interplay of extrinsic barriers and teacher attitudes," *Computers & Education*, vol. 69, pp. 310–319, 2013.

[11] S. Serholt *et al.*, "Teachers' views on the use of empathic robotic tutors in the classroom," in *Proc. of the 23rd IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2014, pp. 955–960.

[12] J. Kory Westlund *et al.*, "Lessons From Teachers on Performing HRI Studies with Young Children in Schools," in *Proc. of the 11th ACM/IEEE Int. Conf. on HRI*. IEEE Press, 2016, pp. 383–390.

[13] E. Lee *et al.*, "Elementary and middle school teachers', students' and parents' perception of robot-aided education in Korea," in *Proc. of the World Conf. on Educational Multimedia, Hypermedia and Telecommunications*, 2008, pp. 175–183.

[14] S. Serholt *et al.*, "The case of classroom robots: teachers' deliberations on the ethical tensions," *AI & Society*, pp. 1–19, 2016. [Online]. Available: http://dx.doi.org/10.1007/s00146-016-0667-2

[15] M. Fridin and M. Belokopytov, "Acceptance of socially assistive humanoid robot by preschool and elementary school teachers," *Computers in Human Behavior*, vol. 33, pp. 23–31, 2014.

[16] S. Y. Okita and A. Jamalian, "Current challenges in integrating educational technology into elementary and middle school mathematics education," *Journal of Mathematics Education at Teachers College*, vol. 2, no. 2, pp. 49–58, 2011.

[17] M. E. Karim *et al.*, "A review: Can robots reshape k-12 stem education?" in *Proc. of the 2015 IEEE Int. Workshop on Advanced Robotics and its SOcial impacts*, no. EPFL-CONF-209219, 2015.

[18] P. Dillenbourg *et al.*, "Classroom orchestration: The third circle of usability," in *Proc. of the 9th Int. Conf. on Computer-Supported Collaborate Learning*, vol. 1, 2011, pp. 510–517.

[19] O. S. Jarrett, "Science interest and confidence among preservice elementary teachers," *Journal of Elementary Science Education*, vol. 11, no. 1, pp. 49–59, 1999.

[20] E. A. Skinner and M. J. Belmont, "Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year." *Journal of Educational Psychology*, vol. 85, no. 4, p. 571, 1993.

[21] U. Schiefele, "Interest, learning, and motivation," *Educational Psychologist*, vol. 26, no. 3-4, pp. 299–323, 1991.

[22] O. Köller *et al.*, "Does interest matter? the relationship between academic interest and achievement in mathematics," *Journal for Research in Mathematics Education*, pp. 448–470, 2001.

[23] G. Gordon *et al.*, "Can Children Catch Curiosity from a Social Robot?" in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI*. ACM, 2015, pp. 91–98.

Towards "Machine-Learnable" Child-Robot Interactions: the PInSoRo Dataset

Séverin Lemaignan[1], James Kennedy[1], Paul Baxter[2] and Tony Belpaeme[1]

***Abstract*— Child-robot interactions are increasingly being explored in domains which require longer-term application, such as healthcare and education. In order for a robot to behave in an appropriate manner over longer timescales, its behaviours should be coterminous with that of the interacting children. Generating such sustained and engaging social behaviours is an on-going research challenge, and we argue here that the recent progress of deep machine learning opens new perspectives that the HRI community should embrace. As an initial step in that direction, we propose the creation of a large open dataset of child-robot social interactions. We detail our proposed methodology for data acquisition: children interact with a robot *puppeted* by an expert adult during a range of playful face-to-face social tasks. By doing so, we seek to capture a rich set of human-like behaviours occurring in natural social interactions, that are explicitly mapped to the robot's embodiment and affordances.**

I. Machine Learning: the Next Horizon for Social Robots?

While the family of *recurrent neural networks* have repeatedly made the headlines over the last few years with impressive results, notably in image classification, image labelling and automatic translation, they have been largely ignored in many other fields so far as they are perceived to require very large datasets (hundreds of thousands to millions of observations) to actually build up useful capabilities. Even though neural networks have demonstrated compelling results in open-ended, under-defined tasks like image labelling, they did not stand out as attractive approaches to problems involving high dimensions with relatively small datasets available – like human-robot social interactions.

Besides, if one considers "social interactions" to also entail joint behavioural dynamics, and therefore, some sort of temporal modeling, neural networks look even less enticing as time is notably absent from most of the tasks which neural networks have been successful at.

In 2015, the Google DeepMind team demonstrated how a convolutional recurrent neural network could learn to play the game Break-Out (amongst 48 other Atari games) by only *looking* at the gaming console screen [1]. This result represents a major milestone: they show that a relatively small sample size (about 500 games) is sufficient for an artificial agent to not only learn how to play (which requires an implicit model of time to adequately move the Break-Out paddle), but to also create gaming strategies that *look like* they would necessitate planning (the system first breaks bricks on one side to eventually get the ball to break-out and reach the area *above* the remaining bricks, therefore ensuring rapid progress in the game). We argue that the complexity of mechanisms that such a neural network has been able to quickly uncover and model should invite our community to question its applicability to human-robot interactions (HRI) in general, and sustained, natural child-robot interactions in particular.

However, the lack of a widespread HRI dataset suitable for the training of neural networks is a critical obstacle to this initial exploration. Therefore, as a first step, we propose a design for such a dataset, as well as a procedure to acquire it. We hope that discussions during the workshop may help in further refining this proposal.

II. Machine learning and social behaviour

Using interaction datasets to teach robots how to socially behave has been previously explored, and can be considered as an extension of the traditional learning from demonstration (LfD) paradigms to social interactions (for instance [2], [3]). Previous examples have generally focused on low-level recognition or generation of short, self-standing behaviours, including social gestures [4] and gazing behaviours [5].

Based on a human-human interaction dataset, Liu *et al.* [6] have investigated machine learning approaches to learn longer interaction sequences. Using unsupervised learning, they train a robot to act as a shop-keeper, generating both speech and socially acceptable motions. Their approach remains task-specific, and while they report only limited success, they emphasise the "life-likeness" of the generated behaviours.

Kim *et al.* [7] highlight that applying deep learning to visual scene information in an HRI scenario was successful, but that generating behaviours for the robot to be able to act in a dynamic and uncertain environment remains a challenge.

These examples show the burgeoning interest of our community for the automatic learning of social interactions, but also highlight the lack of structure of these research efforts, as further illustrated by the quasi-absence of public and large datasets of human-robot interactions. To our best knowledge, only the H^3R Explanation Corpus [8] and the Vernissage Corpus [9] have been published to date. The H^3R Explanation Corpus is a human-human and human-robot

This work has been partially supported by the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227), the EU FP7 DREAM project (grant 611391), and the EU H2020 L2TOR project (grant 688014).

[1]Séverin Lemaignan, James Kennedy and Tony Belpaeme are with the Centre for Robotics and Neural Systems, Plymouth University, U.K. firstname.surname@plymouth.ac.uk

[2]Paul Baxter is with the Lincoln Centre for Autonomous Systems, School of Computer Science, University of Lincoln, U.K. pbaxter@lincoln.ac.uk

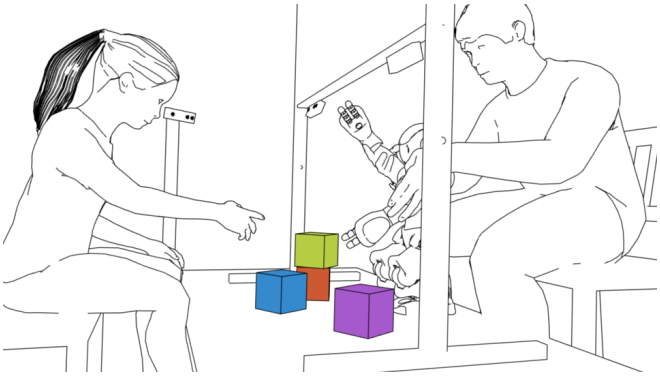


Fig. 1. The acquisition setup: a child interacts with a robot in a range of interactive tasks. The robot is physically guided by an adult expert. We record, in a synchronised manner, the full joint-states of the robots, the RGB and depth video stream from three perspectives (global scene and each of the participant faces), and the sounds (notably, the verbal interactions between the participants).

dataset focusing on a "assembly/disassembly explanation" task and includes physiological signals (22 human-robot interactions), but is not publicly available. the Vernissage Corpus includes one museum guide robot interacting with two people (13 interactions in total), with recordings and annotations of poses and speech audio (stated to be publicly available). Both these corpora are however too small for machine-learning applications.

III. The Plymouth Interacting Social Robots Dataset (PInSoRo)

A. High-Level Aims

The Plymouth Interacting Social Robots (PInSoRo) Dataset is intended to be a novel dataset of human-guided social interactions between children and robots. Once created, we plan to make it freely available to any interested researcher.

This dataset aims to provide a large record of social child-robot interactions that are *natural*: we aim to acquire robot behaviours through corresponding human social behaviour. To this end, we propose that an expert adult will *puppet* a passive robot (Fig. 1). As such, the gestures, expressions and dynamics of the interaction are defined and acted by a human, but as he/she uses the robot body to actually perform the actions, the motions are implicitly constrained by (and thus reflect) the robot embodiment and affordances.

The interactions are supported by a range of short social tasks (described in Section III-B). Critically we propose to limit these tasks to *face-to-face* social interactions, either dyadic or triadic. This constrains the dataset to a more tractable domain, and should ensure technical feasibility. The tasks have to fulfil several key requirements:

- be *fundamentally social*, *i.e.* these tasks would make little or no sense for an agent alone;
- foster rich *multi-modal interaction*: simultaneous speech, gesture, and gaze behaviours are to be observed;
- exhibit *non-trivial dynamics*, such as implicit turn-taking;
- should cover a *broad range of interaction contexts* and situations.

While the tasks will initially be short (in order to acquire a diverse enough dataset), we believe that the captured social behaviours could also be used to inform long-term child-robot interaction. Indeed, naturalistic, rich and socially-oriented multimodal behaviour (beyond simple stereotyped and reactive behaviour) sets the expectation in the human that long-term interactions and social presence [10] can be supported by the robot. Furthermore, we expect such a dataset to allow researchers to uncover several implicit and/or micro-behaviours that, while essential for long-term natural interactions, are difficult to explicitly characterise, and therefore difficult to implement.

B. Tasks

We suggest an initial set of four tasks, lasting about 10 minutes each. They involve collaborative manipulation of simple objects (such as toy cubes), (acted) storytelling, and dialogue-based social gaming. The tasks are intended to be sufficiently different from one another in order to collect a variety of different behaviours, and to minimise task-dependency of the behaviours eventually learnt from the dataset. Physical manipulation of objects across the tasks is limited by the Aldebaran Nao grasping capabilities; the tasks are designed with this in mind, *e.g.* pushing objects away or to the side is possible, whereas pulling them is more difficult.

The tasks are also designed to be playful and engaging, and are derived from classic childrens' games and activities (they are directly inspired by tasks used in other child-robot interaction work, such as [11]). They are thus expected to elicit social interactions that are particularly relevant to child-robot interaction.

a) Task 1: Spatial reasoning: In this task, one partner (child or robot) has a "completed" model made from shapes. Their role is to explain to the other partner how to arrange an identical set of shapes in order to re-create the completed model. The partner with the completed model is not allowed to directly touch the shapes. This task is intended to encourage verbal communication and deictic as well as iconic gestures. It is possible to tune the difficulty of the task through, for example, providing multiple pieces with the same colour, or shape. Similar spatial tasks have been used in other HRI experiments both with adults [12] and children [13].

b) Task 2: Storytelling: The second task revolves around storytelling. To provide a context and collaborative element to the storytelling, "Story Cubes" are incorporated into the task. These cubes are like dice, but with pictures in place of numbers; the pictures serve to guide the story. The two partners are asked to invent a story together, and they take turns in throwing one (large, custom-made) die, arranging the new picture into the story line, and proceeding to tell, and act out, the unfolding story. This task is expected to primarily generate verbal interaction, accompanied by iconic gestures.

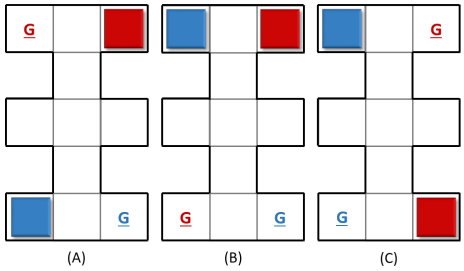


Fig. 2. A sokoban-inspired task requiring collaboration to complete given limitations in robot manual dexterity: the robots face each other across the long edge of each puzzle. Each object (red/blue square) must be pushed to its own goal (red/blue G), in three example levels of difficulty: (A) red and blue objects each simply pushed by one individual, both interactants required, but no explicit collaboration; (B) again a single object requires only a single interactant to manipulate, but some coordination is required due to shared path; (C) each object requires both interactants to manipulate, as well as coordination due to joint path.

c) Task 3: Collaborative strategising: The third proposed task is inspired by the *Sokoban* game (Fig. 2): the two partners must correctly move a set of cubes to locations within a 2D playground by only *pushing* the cubes. Due to the physical setup of the interaction (Fig. 1), the robots are essentially limited to pushing *away* the cubes, transforming the game into a necessarily collaborative activity.

d) Task 4: Party game "Taboo": The fourth proposed task involves triads in a social party game chosen not to require specific gesturing. One such game is "Taboo", a game where one must get others to guess a word without using the word itself. As the game relies only on verbal interaction, we expect all the gestures and gaze behaviour performed by the players to be social backchannel communication, and therefore of direct relevance for the dataset. Using triads is also expected to elicit a richer set of social situations. We expect it to prevent the overfitting of the model to the specific features of dyadic social interactions.

C. Methodology

The envisioned dataset would be comprised of a large number (> 50) of about 30 minutes long recordings of interactions between one child and one puppet-robot, guided by an experimenter (Fig. 1). The pair would be invited to play one or several of the proposed tasks (to be defined after initial pilots). The children would be between 8 and 14 years old. A possibly narrower age range is to be specified once the tasks are precisely defined to ensure the tasks are suitable and engaging for the target age group. Children would typically be recruited from local schools.

We propose to use a Nao robot, and to record the full joint-state of the robot over time. The robot is mostly passive: the feet are firmly fixed on the support table, and all other degrees of freedom, except for the head, are free. The head is externally controlled so that the robot gaze follows the gaze of its human puppeteer in real-time.

The choice of the Nao robot is guided by its small size, making it suitable for puppeting, and its prevalence in the HRI community, resulting in a dataset relevant for a broader academic audience. Also, since Nao is a relatively high degrees-of-freedom (DoF) robot (25 DoFs in total, 5 DoFs per arm), it mimics human kinematics reasonably well. As the motions are recorded in joint space, the dataset can be mapped to other robotic embodiments with similarly configured degrees-of-freedom.

D. Recorded Data

The dataset would comprise the following raw data:

- full 30Hz 25 DoF joint-state of the Nao robot,
- RGB + depth video stream of the scene (see Fig. 1),
- RGB + depth video stream from the child, as seen by the robot,
- speech recording.

Recorded in a fully synchronised manner, these data streams are intended to represent a useful input for many machine-learning techniques. They provide a rich dataset for a range of domains related to social child-robot interaction: from analysis of behavioural alignment between partners (via metrics like the recently proposed *Individual Motor Signature* [14]), to modeling of the dynamics of turn-taking, to the uncovering of implicit in-the-moment synchronisation mechanisms.

This would be complemented by higher-level, post-processed data:

- 68 face landmarks on the child's face, providing options for further facial analysis (like emotion recognition),
- child's skeleton extraction,
- the gaze localisation of each of the participants,
- the 3D localisation of all physical actors (child, all robot parts, cameras, table, manipulated objects),
- the verbal interaction transcripts (automatic transcript with manual verification and correction).

All these sources would be acquired via the ROS middleware (which provides the required mechanism for time synchronisation between the sources) and stored as *ROS bag* files, making it simple to replay the interactions.

As this dataset would contain sensitive data involving children, strict and specific guidelines to ensure the ethical handling of the dataset will be issued before effectively sharing any data.

IV. Discussion

A. Envisioned Applications

The recent advances in machine-learning described in the introduction raise the question of its applicability to the key challenges of artificial intelligence for robotics. Social HRI is a particularly difficult field as it encompasses a large range of cognitive skills in an intricate manner. Application domains of social HRI are typically under-defined, highly dynamic and difficult to predict.

From the data collected, a starting point for machine learning could entail a probabilistic model for reactive behaviours in a given task, *i.e.* finding for each "social cue" the possible set of responses and their probabilities. This could be made generative by using the probability distribution to

seed a roulette-wheel action selection mechanism, effectively creating a probabilistic reactive controller. Whilst simplistic, this is an illustrative example of how the data may be used.

As suggested in the introduction, we also believe that such a dataset could be used to train deep neural networks. While the proposed dataset is very likely not comprehensive enough to train a neural network into an autonomous interactive system, it may be sufficiently rich to train interesting hidden units whose activations would be conditional on specific social situations. For instance, one could imagine that an adequately configured network would generate hidden units able to activate on joint gaze, or on deictic gestures. It must be emphasised that such findings are entirely hypothetical, and we only conjecture them here.

B. Possible Methodological Alternative

Several methodological issues that may impact on the quality of the interaction, the data collection, and the generalisability of results have been anticipated. As the puppeteer behaviours are bound to the embodiment of the robot, it may be that this manipulation inhibits the production of natural behaviours. A small-scale pilot will be used to explore whether or not the puppetted behaviours of the robot inhibit natural interactions with the children.

Besides, one drawback of the proposed acquisition methodology is that the puppeteer remains partially visible to the child (the hands, legs, torso are visible), which may impact the clarity of the interaction (is the child interacting with the robot or with the human behind it?). An alternative acquisition procedure is considered where the puppeteer would remotely control the robot from a different room, using Kinect-based skeleton tracking for the posture control, a head-mounted device for immersive remote vision, and a headset for remote audio. While this adds significant complexity to the acquisition procedure and increases the level of dexterity a task may require, it would provide a cleaner interaction context.

While the tasks have been designed to collect a variety of social behaviours and interaction dynamics, it may be that they are still too similar for any subsequent machine learning to acquire adequately general (*i.e.* not task-specific) behaviours for broader use. Similarly, the use of a single robot may prevent generalisation to other robotic platforms. However, it is not possible to know until algorithms have been applied and tested.

C. Long-Term Considerations

If *useful* social behaviours can be learnt from the initial dataset collected, then this would warrant further collection and exploration of the technique. Transfer to adult-adult pairs could be conducted (possibly with modification of the tasks). Child pairs performing the tasks without the robot could be used to further update behavioural models, as could human behaviours in response to learned robot models, thus providing longer-term adaptivity of behaviour.

Whilst we must acknowledge that the task-centred interactions we propose as part of the PInSoRo dataset are relatively short-term, we do argue that they are capable of simultaneously capturing a range of subtle and complex naturalistic behaviours across a range of different modalities. This type of rich behaviour (by going beyond simple stereotyped and reactive behaviour) supports the expectation in the human that they are interacting with a truly socially competent agent, thus providing the conditions in which long-term child-robot interactions could take place. The application of machine learning algorithms (particularly "deep" methods) provide an opportunity to automatically datamine the solutions to this vastly complex problem that may not be possible with hand-coded systems. Whilst this methodology may yet prove to not be *sufficient* for a complete solution, we propose that the PInSoRo dataset (and others that may follow) establishes a *necessary* foundation for the creation of socially-competent robots over long-term interactions.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] C. L. Nehaniv and K. Dautenhahn, *Imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions*. Cambridge University Press, 2007.

[3] Y. Mohammad and T. Nishida, "Interaction learning through imitation," in *Data Mining for Social Robotics*. Springer, 2015, pp. 255–273.

[4] Y. Nagai, "Learning to comprehend deictic gestures in robots and human infants," in *Proc. of the 14th IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2005, pp. 217–222.

[5] S. Calinon and A. Billard, "Teaching a humanoid robot to recognize and reproduce social cues," in *Proc. of the 15th IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2006, pp. 346–351.

[6] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "How to train your robot - teaching service robots to reproduce human social behavior," in *Proc. of the 23rd IEEE Int. Symp. on Robot and Human Interactive Communication*, 2014, pp. 961–968.

[7] K.-M. Kim, C.-J. Nan, J.-W. Ha, Y.-J. Heo, and B.-T. Zhang, "Poro-robot: A deep learning robot that plays video q&a games," in *Proc. of the AAAI 2015 Fall Symposium on AI for Human-Robot Interaction*, 2015.

[8] Y. Mohammad, Y. Xu, K. Matsumura, and T. Nishida, "The H^3R Explanation Corpus human-human and base human-robot interaction dataset," in *Proc. of the Int. Conf. on Intelligent Sensors, Sensor Networks and Information Processing*. IEEE, 2008, pp. 201–206.

[9] D. B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J.-M. Odobez, S. Wrede *et al.*, "The vernissage corpus: A conversational human-robot-interaction dataset," in *Proc. of the 8th ACM/IEEE Int. Conf. on Human-Robot Interaction*. IEEE Press, 2013, pp. 149–150.

[10] I. Leite, C. Martinho, A. Pereira, and A. Paiva, "As time goes by: Long-term evaluation of social presence in robotic companions," in *Proc. of the 18th IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2009, pp. 669–674.

[11] T. Belpaeme, J. Kennedy, P. Baxter, P. Vogt, E. J. Krahmer, S. Kopp *et al.*, "L2TOR - Second Language Learning Tutoring using Social Robots," in *Proc. of the First Int. Workshop on Educational Robots at the 2015 Int. Conf. on Social Robotics*, 2015.

[12] A. Sauppé and B. Mutlu, "Effective task training strategies for instructional robots," in *Proc. of the 10th Annual Robotics: Science and Systems Conference*, 2014.

[13] C. Zaga, M. Lohse, K. P. Truong, and V. Evers, "The effect of a robot's social character on children's task engagement: Peer versus tutor," in *Proc. of the 2015 Int. Conf. on Social Robotics*. Springer, 2015, pp. 704–713.

[14] P. Słowiński, C. Zhai, F. Alderisio, R. Salesse, M. Gueugnon, L. Marin *et al.*, "Dynamic similarity promotes interpersonal coordination in joint action," *Journal of The Royal Society Interface*, vol. 13, no. 116, 2016.

A Cautionary Note on Personality (Extroversion) Assessments in Child-Robot Interaction Studies

Paul Baxter, Tony Belpaeme
Centre for Robotics and Neural Systems
Plymouth University, U.K.
{paul.baxter, tony.belpaeme}@plymouth.ac.uk

***Abstract*—The relationship between personality and social human-robot interaction is a topic of increasing interest. There are further some indications from the literature that there is an association between personality dimensions and various aspects of educational behaviour and performance. This brief contribution seeks to explore the single personality dimension of extroversion/introversion: specifically, how children rate themselves with a validated questionnaire in comparison to how teachers rate them using a relative scale. In an exploratory study conducted in a primary school, we find a non-significant association between these two ratings. We suggest that this mismatch is related to the context in which the respective ratings were made. In order to facilitate generalisation of personality-related results across studies, we propose two general reporting recommendations. Based on our results, we suggest that the application of personality assessments in a child-robot interaction context may be more complex than initially envisaged, with some dependence on context.**

I. Introduction

The role of personality in human-robot interaction is becoming of greater interest in the field, as attempts are made to increase the adaptability and personalisation of the robots. For example, preference has been found in a rehabilitation context for a robot that matches one's own personality [1]. Similarly, in children, robots that take into account the personality of the interacting child (e.g. if shy) can adapt its behaviour accordingly to promote interaction [2]. In our research, we are generally interested in having robots adapt to children within interactions in order to facilitate some outcome such as learning or behaviour change, e.g. [3]. As a trait upon which adaptation can be based, personality is therefore of interest.

There are a number of problems with administering of lengthy questionnaires to children, particularly those related to personality assessments. Primarily, these include the level of concentration required for completion, and the conceptual level of the questions (specifically if abstract, or relating to life experiences that may not be typical for children). For this reason, there have been a wide range of development and validation efforts to produce short-form questionnaires in a range of languages. Part of this validation process frequently involves examining the convergence of personality ratings between parents, teachers and the children themselves, with high agreement being used to support validity, e.g. [4]. One such effort is an abbreviated Junior Eysenck Personality Questionnaire [5], which attempts to characterise four dimensions of extroversion, neuroticism, psychoticism, and (questionnaire) reliability using 24 questions. This was validated with children in the age range 13–15, although related work showed valid application to children of a slightly younger age [6]. In the present study, we employed a short-form version of a five-factor model questionnaire that has been validated with children [4]: the BFQ-C.

We focus specifically on the single dimension of extroversion. Prior work has, for example, suggested that extroversion is positively associated with verbal-imagery-based learning in children [7], and with help-seeking behaviours (self-regulated learning) in adults [8]. These make it a dimension of interest to our educational context. Extroversion is also suitable as a characteristic of interest since it is a dimension (extroversion vs introversion) that appears in a range of human personality theories (for example both the Eysenck and five-factor 'big 5' models).

In this study, we examine the relationship between self-rated scores of extroversion with teacher ratings of relative extroversion. As a secondary consideration, we also consider the possible relationship to learning outcomes in a subsequent collaborative learning task with a social robot, although this is not the focus of this paper. The work described here accords with our wider goals of ethologically-appropriate and valid empirical investigations for child-robot interaction in educational contexts [9]. First we introduce the exploratory study (section II), before interpreting the results (section II-B) as suggesting that care must be taken in considering the context of the personality assessment (section III).

II. Exploratory Study

As an exploratory study, we do not propose hypotheses. However, from the discussion above, we may venture the predictions that the child self-ratings of extroversion and the teacher-ratings of the same will be positively associated (reflecting that the teachers know the children), and that there will be a positive association between ratings of extroversion and learning outcome. In the following, we assess whether the data provide any support for these predictions.

A. Setup and Method

The study was conducted in two primary schools in the U.K. 38 children, aged 7–8 years old took part (22 boys, 16 girls). The study was run in accordance with a protocol approved by the Plymouth University Faculty of Science and

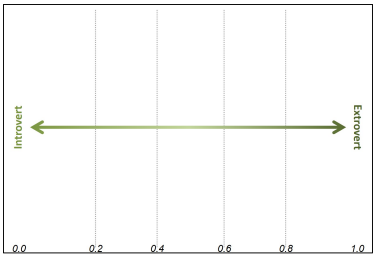


Fig. 1. The teacher scale for relative introversion/extroversion child ratings. Teachers were instructed to write the names of the children on the sheet. Numbers in italics (not displayed to the teachers) indicate coding of the position of the names on the sheet: names on (or next to) the dotted were assigned the score shown; numbers in the space between dotted lines were assigned an intermediate score (e.g. 0.3, 0.5, ...).

Technology ethics board. An opt-out consent was obtained from all parents/guardians of the children, with separate opt-in consent for image/video recording (not used in the present paper). All children were permitted to withdraw from the study at any point upon request. The experiment took place towards the end of the school year, meaning that the teachers had spent at least the majority of a school academic year with the children.

The visit to each school began in the morning: after initial attendance check, the experimenters were introduced to the class. They informed the children of the purpose of the visit: to play a sorting game with the robot and to fill in some questionnaires (both knowledge pre/post tests and the personality questionnaire). The extroversion scale questionnaire was then administered to the children as a group in the classroom, led by the teacher: independent completion was instructed (and enforced) by the teacher (i.e. prevention of copying).

Separately, the teacher was briefed on their rating of the children's extroversion. On the single dimension of extroverted to introverted, represented on a single sheet of paper (figure 1), the teachers were asked to place the children in their class in relation to one another, based on a similar rating scheme as used in [10]. This scale was therefore an explicitly subjective and relative rather than a subjective and independent measure of this personality characteristic. The intention was to examine the correlation between teacher-ratings and self-ratings rather than a direct comparison of scores.

Through the rest of the day, the children were brought one-by-one into a separate room in the school, where they completed a pre-knowledge test on carbohydrates, engaged in a sorting task with a Nao humanoid robot (Aldebaran Robotics) on the topic of carbohydrates for five minutes, and then completed a post-knowledge test (different pre and post tests, counterbalanced between individuals, not containing images used on the interaction).

B. Results

A qualitative inspection of the data does not suggest any strong relationships between child self-rating of extroversion, the teacher-rating of the same, and learning outcome (figure 2). The correlation between learning outcome, as measured by pre- to post-test score change, and both self-rating (r=0.030, p=.857, n=38) and teacher-rating (r=0.029, p=.863, n=38) is not significant, with very low effect sizes. Due to the incidental nature of learning outcome for the present contribution, we do not consider it further, other than to note this lack of significant association.

TABLE I
CORRELATION COEFFICIENTS PER QUESTION (n=38 FOR ALL), BETWEEN INDIVIDUAL RATINGS AND OVERALL SELF-RATING, AND OVERALL TEACHER-RATINGS. SIGNIFICANT CORRELATIONS ($\alpha = .05$) ARE HIGHLIGHTED IN GREEN; ITALISISED VALUES HAVE MARGINAL p-VALUES (IN RANGE $0.05<p<0.06$).

Question	Pearson Correlation	
	Self-rating	Teacher-rating
Q1	0.4994	0.1207
Q2	0.1845	-0.2115
Q3	0.4734	*0.3090*
Q4	0.4141	-0.1541
Q5	*0.3112*	0.0076
Q6	0.3872	0.1234
Q7	*0.3187*	0.2414
Q8	0.5706	-0.0226
Q9	0.5675	0.1494
Q10	0.1577	0.1348
Q11	0.4890	0.0900
Q12	0.5463	-0.1016
Q13	0.3937	0.1888

Of perhaps more unexpected nature is the low (non-significant) correlation between the teacher-rating and the child self-rating (r=0.142, p=.142, n=38, figure 2(a)). This indicates that there was weak agreement between the children and their teachers, despite spending extended periods of time with each other (i.e. the school days).

Examining the correlations between the data obtained on a single question basis further supports the observation that there is at best only a weak link between the self-ratings and the teacher-ratings. Firstly, as would be expected, there is generally a high number of positive correlations between the individual question responses and the overall self-rating (table I). Secondly, however, this positive relationship is not reflected in the correlation of self-ratings to the overall teacher-ratings. Only for one question (Q3: "I like to move and to do a great deal of activity") is there a moderate positive correlation (though not quite significant, p=0.059, n=38). Interestingly, this positive correlation between physical activity and extroversion has been found in children of this age [11], indicating some (limited) support for the idea that the teachers do have some familiarity with the children, and that there is divergence between the ratings in spite of this.

A further result of interest is related to Q8 ("I like to talk with others"). There is a strong positive correlation between the overall self-rating and the response to this question (r=0.571, $p<.001$, n=38), but there is no correlation between the response to this question and the teacher-rating of extroversion (r=-0.023, p=.893, n=38). Assuming that willingness of children to speak with others is likely to be one of the more apparent characteristics of children to their teachers, this lack of association is perhaps surprising.

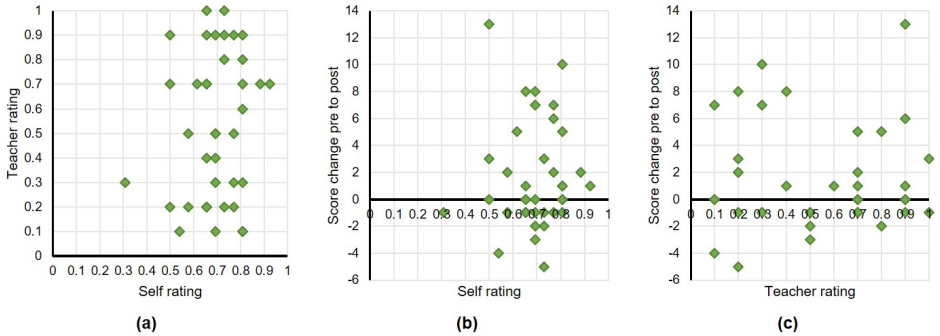


Fig. 2. Raw data scatter plots showing the relationship between the metrics of child self-report extroversion (normalised scale), teacher-reported child extroversion (normalised scale), and learning outcome (*post-test score - pre-test score*): (a) self-rating versus teacher-rating; (b) self-rating versus learning outcome; and (c) teacher-rating versus learning outcome.

III. Discussion

The results of this exploratory study suggest that there is a difference between the way the children see themselves in terms of extroversion and the way their teachers see them. Extroversion in this context seems to be a particularly relevant characteristic to explore in this way given its overt behavioural component. These results are consistent with previous observations that there is little agreement between child self-ratings and teacher-ratings (although parent-ratings fared better) [10], although the present study extends these by significantly extending the number of subjects involved. Indeed, the present results also seem to be in accordance with the results from the questionnaire validation itself, which suggested a non-significant association between teacher ratings and child self-ratings for extroversion, but only for younger children [4].

We, in this and other work, examine interactions between children and robots in school environments. We are therefore interested in characterising various aspects of this context in particular.

One consideration having an effect on these results may be the environment in which the study was conducted, and the relationship between this and the teacher as involved observer. The teacher interacts with the children only during the school day (with the type of educational environment itself providing potential biases of child behaviours and teacher interpretations thereof), and would typically not do so outside of the context of school. The children are naturally not constrained by school alone, and as such will have a broader experience upon which they base their personality self-assessments. It is thus perhaps not surprising that there is an apparent mismatch between the children's perception of themselves (albeit on only one personality sub-scale) and that of their teachers. The question then becomes, which assessment (child or teacher) is more relevant to school performance? This is only speculation, but the results provide an basis for further empirical exploration.

There are a number of issues, both general and specific, with the present study. Firstly, we were limited in our examination of only one sub-scale of personality as characterised by the 5-factor model. Furthermore, while a questionnaire validated with children was used, the three-item scale we employed (for reasons of clarity for the children) is relatively coarse, thus limiting the resolution of the measure. Nevertheless, the wide range of responses obtained (see Appendix) suggests that inter-personal variability was still discernible. Secondly, we were comparing self-ratings from the questionnaire with relational ratings from a third party (the teacher). Being relational, this explicitly rated the children with respect to one another: our prediction that it is reasonable to examine the association of the two measures is clearly not borne out by the results. While we interpret this as a context effect, there is naturally the possibility that it is our comparative measures approach that is flawed. The mis-matching results nevertheless remain to be explained, and thus still in our view constitute a reason to be wary of the self-rating (or other-rating) measure alone. Thirdly, compared with personality questionnaire validation exercises (with participant numbers typically in the multiple hundreds), our sample size (n=38) is relatively small. Given that at large sample sizes moderate to small correlation coefficients can become statistically significant, it is possible that a more extended study would find that our results were also significant. However, the low effect size (r=0.142) still suggests the lack of a straightforward positive association between self- and teacher-ratings. Finally (and in general), there are also a number of issues related to the administration of questionnaires to children, e.g. [12], as a result of effects such as social desirability, thus calling into question the reliability of such methods. While validation of the questionnaire with the appropriate subject group (i.e. children in this case) can mitigate this effect, it is necessary to remain cautious.

While these issues naturally reduce the potential power of the results obtained, we believe that there are still a number of pertinent points that are raised by this study. Generally, and this is of course an important consideration for any empirical investigation, how can we be sure that we are measuring what we intend to measure? For our particular case, this was the (relative) extroversion of the children who took part in the study: the issue is whether the questionnaire used was adequate given the context in which it was completed, and whether the teacher-ratings (using the relative rating method, figure 1)

can give a 'true' reflection of the children's extroversion in the context of the classroom at least – and indeed whether this could be different from a rating in a different context. Validation of the questionnaire [4] suggests that it is a reliable measure, but does this extend across all contexts? We assumed that rating extroversion would be reasonably assessed by the teachers given that it relates to observable behaviours (as well as attitudes) that would be reasonably expected to be manifested by the children in the classroom environment.

IV. Conclusions

In describing these apparently problematic elements of the present study (section III), we do not seek to suggest that there is no value in exploring personality characteristics and its relationship to behaviours and performance. Instead, the case study presented here suggests that the methodological and reporting standards of such characteristics require clarity, in line with similar suggestions for the field of HRI in general [13]. In order to facilitate this, and to promote generalisation to (and comparison with) other studies, we suggest the following (modest) guidelines:

G1. Identify the source of the personality questionnaire (or other characterisation method) used in terms of the assumed dimensions (e.g. the 'Big 5' or the Eysenck dimensions), and whether it has been validated with the age group (and indeed language) under consideration.

G2. Identify the context in which the children completed the personality characterisation, and indicate possible influencing factors (e.g. completed in the presence of teachers/friends/parents, at home/school, in group/individually).

These guidelines are not particularly novel, and do in fact simply promote the complete reporting of measures and possible confounds. However, through our exploratory study we hope to have demonstrated that an apparent straightforward characterisation of one aspect of personality involves a number of complicating factors that should themselves be characterised. If the results we obtained were anomalous in some way, we hope that by reporting these potential confounds other researchers can build on them, by either accounting for the effect, or discounting it through further investigation. At the present time however, in our discussion of the results we highlight the possibility that child self-ratings of extroversion may be unreliable for child-robot interaction studies, whether this is due to inherent age-related unreliability, environment context effects, or others.

We do not suggest that we have a solution to the apparent issues described in this paper, particularly the mismatch between child- and teacher-ratings of extroversion, although we do venture some ideas for why this occurred. Indeed, we recognise a number of limitations in the study that prevent the formulation of a solution. Nevertheless, the suggestion remains that the application of personality assessments in a child-robot interaction context may be more complex than may be initially envisaged, with some dependence on context. As such, we suggest that the proposed guidelines will at least provide a basis upon which progress can be made.

Acknowledgement

This work was supported by the EU FP7 project DREAM (grant number 611391, http://dream2020.eu), and the H2020 project L2TOR (grant number 688014, http://www.l2tor.eu).

References

[1] A. Tapus, C. Tapus, and M. J. Matarić, "User–robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy," *Intelligent Service Robotics*, vol. 1, no. 2, pp. 169–183, 2008.

[2] K. Abe, et al, "Toward playmate robots that can play with children considering personality," *Proceedings of the second international conference on Human-agent interaction - HAI '14*, pp. 165–168, 2014.

[3] T. Belpaeme, et al, "Multimodal Child-Robot Interaction: Building Social Bonds," *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 33–53, 2012.

[4] C. Barbaranelli, G. V. Caprara, A. Rabasca, and C. Pastorelli, "A questionnaire for measuring the Big Five in late childhood," *Personality and Individual Differences*, vol. 34, no. 4, pp. 645–664, 2003.

[5] L. J. Francis, "The development of an abbreviated form of the revised junior eysenck personality questionnaire (jepqr-a) among 13–15 year olds," *Personality and Individual Differences*, vol. 21, no. 6, pp. 835–844, 1996.

[6] L. J. Francis and E. M. Thomas, "Welsh language adaptation of the short-form junior eysenck personality questionnaire revised (jepqr-s)," *Psychologist in Wales*, vol. 21, pp. 25–32, 2008.

[7] J. Riding and V. Dyer, "The relationship between extraversion and verbal-imagery learning style in twelve-year-old children," *Personality and Individual Differences*, vol. 1, no. 3, pp. 273–279, 1980.

[8] T. Bidjerano and D. Y. Dai, "The relationship between the big-five model of personality and self-regulated learning strategies," *Learning and Individual Differences*, vol. 17, no. 1, pp. 69–81, 2007.

[9] P. Baxter, et al, "The Wider Supportive Role of Social Robots in the Classroom for Teachers," in *1st Int. Workshop on Educational Robotics at the Int. Conf. Social Robotics*, Paris, France, 2015.

[10] S. M. Robben, "It's NAO or Never Facilitate Bonding Between a Child and a Social Robot: Exploring the Possibility of a Robot Adaptive to Personality ." Radboud Universiteit, Nijmegen, The Netherlands, Tech. Rep., 2011.

[11] D. M. Buss, J. H. Block, and J. Block, "Preschool activity level: Personality correlates and developmental implications," *Child Development*, vol. 51, no. 2, pp. 401–408, 1980.

[12] I. Baroni, et al, "What a robotic companion could do for a diabetic child," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RoMAN'14)*. Edinburgh, U.K.: IEEE Press, 2014, pp. 936–941.

[13] P. Baxter, et al, "From Characterising Three Years of HRI to Methodology and Reporting Recommendations," in *alt.HRI at 11th Int. Conf. on Human-Robot Interaction*, vol. in press. Christchurch, NZ: ACM/IEEE, 2016.

Appendix: Adapted Extroversion Questionnaire

The adapted child-personality questionnaire (BFQ-C; Extroversion scale) used is as shown below [4]. Each Likert scale question had 3 possible responses: [*Almost Never, Sometimes, Almost Always*]. Answers were scored from 1 to 3, respectively, with all responses scored positively. Maximum range of possible responses: [*13, 39*]. Actual range of responses recorded: [*21, 37*], m=31, sd=3.137, n=38.

Q1) I like to meet with other people.
Q2) I like to compete with others.
Q3) I like to move and to do a great deal of activity
Q4) I like to be with others.
Q5) I can easily say to others what I think.
Q6) I say what I think.
Q7) I do something not to get bored.
Q8) I like to talk with others.
Q9) I am able to convince someone of what I think.
Q10) When I speak, the others listen to me and do what I say.
Q11) I like to joke.
Q12) I easily make friends.
Q13) I am happy and lively.

2016

From Characterising Three Years of HRI to Methodology and Reporting Recommendations

Baxter, P

10.1109/HRI.2016.7451777

Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference (alt.HRI)

From Characterising Three Years of HRI to Methodology and Reporting Recommendations

Paul Baxter, James Kennedy, Emmanuel Senft, Séverin Lemaignan, Tony Belpaeme
Centre for Robotics and Neural Systems, The Cognition Institute
Plymouth University, Plymouth, U.K.
Email: {paul.baxter,...,tony.belpaeme}@plymouth.ac.uk

***Abstract*—Human-Robot Interaction (HRI) research requires the integration and cooperation of multiple disciplines, technical and social, in order to make progress. In many cases using different motivations, each of these disciplines bring with them different assumptions and methodologies. We assess recent trends in the field of HRI by examining publications in the HRI conference over the past three years (over 100 full papers), and characterise them according to 14 categories. We focus primarily on aspects of methodology. From this, a series of practical recommendations based on rigorous guidelines from other research fields that have not yet become common practice in HRI are proposed. Furthermore, we explore the primary implications of the observed recent trends for the field more generally, in terms of both methodology and research directions. We propose that the interdisciplinary nature of HRI must be maintained, but that a common methodological approach provides a much needed frame of reference to facilitate rigorous future progress.**

***Index Terms*—Challenges; Human-Robot Interaction; Methodology; Recommendations; Research Methods**

I. Introduction

Human-Robot Interaction as a research field lies at the confluence of multiple disciplines, each with their own goals, assumptions, methodologies and techniques (figure 1). As a result, it provides a rich environment for a variety of research questions and empirical investigations. However, this inherent strength brings with it shortfalls in terms of mismatches between disciplines that should be accounted for. In this paper, we provide an overview of the current state of the field of Human-Robot Interaction through the prism of the ACM/IEEE HRI conference, and on this basis provide a set of guiding principles and technical recommendations that will help to consolidate the progress made thus far, and provide a platform for future contributions. In doing so, we seek to promote introspection in the community to provoke discussion, propagate best practice through our characterisations, and provide a guide to newcomers to the study of HRI – an important aspect given the multidisciplinary nature of the field.

We provide two levels of analysis, from researcher-level to field- and community-level. At the researcher-level, we identify good practice from both within and without the field, and formulate practical recommendations that can be readily applied to ongoing and future research. At the field-level, we consider the broader themes resulting from the inherently interdisciplinary nature of HRI, and how these relate to the methodological and technical challenges faced by researchers. In doing so, we seek to highlight common ground and future directions to provoke discussion in the field and ultimately improve the impact of HRI in terms of both research and applications.

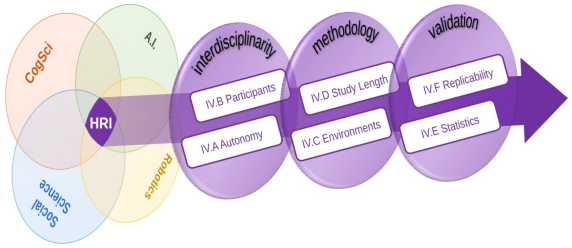


Fig. 1. HRI as a field seeks to integrate knowledge and techniques from multiple disciplines (also including design, psychology, etc), but has its own unique challenges, a number of which we characterise in this paper – numbers correspond to sections in this paper.

We have summarised data from papers presented at the last three HRI conferences (2013, 2014, 2015) to provide recent trends in application and methodology at the primary conference in the field. A total of 101 papers were analysed, with each individual paper classified across 14 categories according to the methods and approaches used within them. This process provides insights into current approaches and emerging trends in the field of HRI.

II. Motivations and Scope

As noted above, each discipline brought into HRI brings with it sets of assumptions and motivations. They may also bring different goals, which may or may not conflict. At the highest level, for instance, we may make a distinction between studies that are theoretically motivated *vs.* application oriented, and between those that are robot centred *vs.* human centred. For example, the use of modelling in cognitive science (where there are increasing numbers of models 'embodied' on robotic platforms) is typically intended to provide an exploration or account of some human-centred phenomenon [1] rather than explicitly seek to improve the robotic agents themselves – although this is on occasion a useful consequence. For robots intended for therapy, e.g. [2], the focus of development is necessarily therapeutic efficacy (i.e. human centred and application oriented) rather than models of robot cognition. In contrast, research to develop physically safe robots to interact with people are more robot-centred and application

oriented emphasising technical contributions, e.g. [3], whereas developmental robotics as applied in human-robot interaction contexts are more robot-centred but theoretically oriented, e.g. [4].

While the presence of this plurality of motivations is not at issue, these differing founding assumptions and intended applications require the use of differing hypotheses and consequently different appropriate methodologies to address them. This is apparent for example when reconsidering the examples from cognitive modelling and therapy: in the former, explicit characterisation of the way a human and robot behave (and possibly how they generate their behaviours) would be necessary, whereas in the latter, a focus would typically be on human behaviour metrics. Whilst such differences do not necessarily result in tension, they can give rise to differing and mismatched expectations between those with different disciplinary backgrounds (as may be expressed in a peer review process for example), typically where the results from one domain are applied to another.

We maintain that this richness is essential for the HRI community, and that it should be preserved. There is a benefit in closer collaboration and the cross-fertilisation of knowledge and methods. One potential means could be to provide a set of benchmarks and target tasks to facilitate comparison between approaches (as with the DARPA or RoboCup@home challenges): a danger of doing so however is the alienation of those parts of the community not engaged in these technical challenges, and the eventual treatment of these benchmarks as ends in their own right, rather than means as originally intended. Therefore, we rather suggest that the provision of a framework to set out common standards and best practice in methodology and reporting centred on the main challenges in the field would encourage and facilitate collaboration and the cross-application of results without bias towards/against any of the disciplines that feed into HRI. To this end, our intention in this paper is to examine and characterise the approaches used in recent HRI conference publications, the challenges that these give rise to, and hence to derive a set of recommendations that can serve as the basis for this common framework.

A reflection of the make up of the conference papers analysed, our perspective in this paper is primarily experimental, irrespective of the actual theme that may have been applied to the paper (e.g. studies, technical advances, design, etc). That is to say, we focus here on the running and reporting of empirical studies rather than theoretical, design or technical contributions in their own right, although we must acknowledge the importance of each of these. Equally, we note that qualitative and ethnographic approaches are fundamentally useful, even if this is not reflected directly in the papers covered in the present review; indeed, the methodological points we discuss below are largely relevant to these approaches in HRI.

In conducting our review exercise in this paper, there are a number of facets of HRI as a field that shaped our decision to focus on recent conference proceedings, with the HRI conference as a particularly important venue, as previously suggested [5]. Since the field is fast paced, with new technological and theoretical developments rapidly shaping the experiments that are run, conference papers provide the most readily and rapidly available results in the peer-reviewed domain, constrasted against the inherently slower publication turn-around of typical journal articles. Our decision to restrict our search to the past three years is similarly intended to explore recent trends given a relatively volatile field.

TABLE I

OVERVIEW OF PAPER AND STUDY TYPES COVERED BY YEAR. NUMBER IN BRACKETS INDICATES FOR EACH CATEGORY THE PERCENTAGE OF PAPERS THAT YEAR. *NHST: Null-Hypothesis Significance Testing*. A 'UNIVERSITY SAMPLE' IS A STUDY WHICH TOOK A SAMPLE OF STUDENTS OR RESEARCH STAFF FROM A UNIVERSITY OR RESEARCH INSTITUTION.

	2013	**2014**	**2015**	**Total**
Number of papers	26	32	43	**101**
With study	25	31	40	**96**
NHST	24 *(96%)*	30 *(97%)*	36 *(90%)*	**90**
University sample	14 *(56%)*	13 *(42%)*	18 *(45%)*	**39**
Lab study	19 *(76%)*	23 *(74%)*	30 *(73%)*	**72**
$>$1 session study	0 *(0%)*	1 *(3%)*	4 *(10%)*	**5**
Uses WoZ	3 *(12%)*	11 *(35%)*	11 *(28%)*	**25**

Through classifying the papers according to the chosen categories, we have identified a number of features of HRI methodology and reporting that warrant consideration, which we have coalesced into six challenges (figure 1 & section III). These challenges are not restricted to any particular disciplinary perspective, but are generally applicable, whilst remaining specific enough to result in practical and actionable recommendations. The aim in doing so is to structure our recommendations so as to provide the foundation for a common frame of reference within which HRI studies with all disciplinary flavours can push the field forward.

III. METHOD

In order to explore the state of the field of HRI, three years of published papers for the Human-Robot Interaction conference were analysed (table I). All 101 full papers from the 2013, 2014 and 2015 proceedings were collated for analysis on the 14 categories shown in table II. All categories were assessed by manually reading the papers and storing the values in a spreadsheet (available at http://goo.gl/PfK1IC).

The categories we chose were ones that were common to all experimental papers, which encompasses the vast majority of papers examined (96 out of 101). They were chosen due to their generality to experimental methodology, being aspects that would be reasonably expected of any study conducted in the field of Human-Robot Interaction. We thus include robot-specific aspects (e.g. nature of control) as well as the standard human-related factors (number of participants, etc), and we suggest that we have included all relevant factors of this nature.

To collect this data, certain definitions were required. Firstly, a lab study is considered to be one in which the participants would have to leave their environment and come to the evaluation location, whereas a non-lab (or 'wild') study is one in which the experimenters go to the participants' environment. Secondly, levels of robot autonomy are described in detail in section IV-A.

TABLE II
OUTLINE OF THE 14 CATEGORIES USED TO CLASSIFY EACH OF THE PAPERS CONSIDERED. *NHST: Null-Hypothesis Significance Testing*.

Category	Classes
Stimuli	Colocated Robot / Non-Colocated Robot / Virtual Robot / Video / Photo / Text / None
Interactive	Yes / No
Robot type/model	Name / N/A
Use of Wizard-of-Oz	Autonomous / Perceptual WoZ / Cognitive WoZ / User Tele-operation / Experimenter Tele-operation / N/A
Occurences of 'wizard'	n / N/A
Study with people	Yes / No
University sample	Yes / No
Mean age participants	Mean / Unstated / Unclear
Conducted in lab setting	Yes / No
Participants per condition	Mean / Unclear / /N/A
Interaction duration (min)	Mean / Unstated / N/A
Interactions per week	n / N/A
Experiment length (weeks)	n / N/A
Use of NHST	Yes / No

The most common unit for each of the relevant categories is used, with translations made if necessary. For papers that present multiple studies, or pilot studies as well as a larger evaluation, only the larger evaluation using a robot, or last study was considered. For interaction durations, if a time range was provided, then the maximum of the range was recorded. Missing data, or cases in which the information was not clear, were annotated in the data collection exercise, with clarification notes appended.

IV. HRI CHARACTERISATION AND RECOMMENDATIONS

Examination of the collected data suggests six broad characteristics that encompass a wide range of non-discipline-specific aspects of HRI research. Roughly following the design process of a system and its subsequent evaluation and reporting, we can consider them to be comprised of (figure 1): robot autonomy and study participants (interdisciplinary aspects), environment and study length (methodological considerations), and statistics reporting and replicability (validation for the community). In the following subsections we provide summary information of the collected data in the 14 identified categories. We note that only 40 of the 96 ($\sim$42%) papers with studies contain all of the information in the 14 categories we examined.

A. Level of Robot Autonomy

We recorded whether or not there was any interaction between the robot (or other stimulus used in an evaluation) and the participants: i.e. those in which the behaviour of the robot is in some way influenced by the behaviour of the interacting human(s). Then, we define several categories in order to assess the levels of autonomy used in HRI studies, shown below, with the results reported in table III. These include a conceptual division in the use of Wizard-of-Oz (WoZ) techniques:

– *Autonomous:* The robot is fully autonomous; minor interventions are still possible, such as starting the system.

TABLE III
AUTONOMY LEVELS ACROSS ALL THREE YEARS OF HRI PUBLICATIONS OF STUDIES, INCLUDING THE IDENTIFICATION OF NUMBER OF *interactive* STUDIES. RELATIONSHIP BETWEEN LEVELS OF AUTONOMY

Autonomy Level	Interactive	Total
Autonomous	38 (*40%*)	46 (*48%*)
Perceptual WoZ	8 (*8%*)	9 (*9%*)
Cognitive WoZ	16 (*17%*)	16 (*17%*)
Participant tele-operation	12 (*13%*)	12 (*13%*)
Experimenter tele-operation	2 (*2%*)	2 (*2%*)
Not Applicable	0 (*0%*)	13 (*14%*)

– *Perceptual WoZ:* The wizard replaces a robotic function (typically a perception capability, such as speech recognition) that could be autonomous (algorithms or tools exist for that function and could have been applied in that context). The function is performed by a wizard for practical reasons (time, difficult technical deployment, computational constraints).

– *Cognitive WoZ:* The wizard replaces cognitive capabilities of the robot, such as deciding what speech to say, what gestures to use, or what actions to take. This can possibly lead the user into ascribing cognitive capabilities onto the robot that do not exist.

– *Participant Tele-operation:* The participant in the study tele-operates the robot as part of the study design, for instance to study shared autonomy.

– *Experimenter Tele-operation:* An experimenter tele-operates the robot as part of the study design, with no intent to deceive participants that the robot has autonomous capabilities (as in the case of WoZ).

– *Not Applicable:* Studies where the autonomy of the robot is not relevant to the procedure, e.g. no robot is present, there is no study, participants watch a video.

In many cases it was difficult to assess the level of autonomy of a robot used in an evaluation. Indeed, 5 papers from 26 utilising a WoZ omit the word 'wizard' altogether. This has previously been raised as an issue in HRI and clear reporting guidelines have already been put forward [6]. Greater adoption of these guidelines would clearly aid the field in understanding the context of the studies conducted.

Note that the level of autonomy, as per our definition, is not to be taken as a proxy for the system (or experiment) *complexity*: some of the systems labeled as autonomous implement simple, fully scripted interactions. On the contrary, some of the wizarded experiments do involve complex autonomous processing for certain parts.

Wizard-of-Oz, as a manipulation technique, is often an experimentally appropriate methodology. A case in point consists in using the robot as a puppet to uncover specific social human behaviours when confronted with a machine (which is typical for the *human centred, theory focused* research line introduced in section II).

When employed, Wizard-of-Oz necessitates special care: since the interaction becomes partially (or in some cases, entirely) a human-human interaction, mediated by a 'mechanical puppet', the researchers need to ensure replicability of the wizarded behaviours between participants, and be careful not to

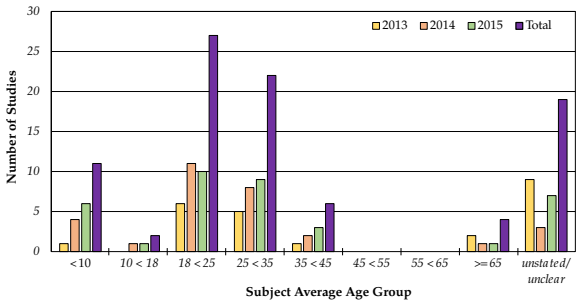


Fig. 2. Histogram of the average age of evaluation participants by age and total from the last 3 years of HRI conference publications. There is a clear peak for the age of student-based samples.

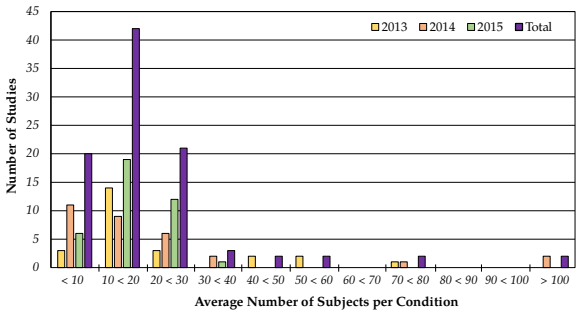


Fig. 3. Histogram of the average number of participants per condition of evaluations from the last 3 years of HRI conference publications. The majority of conditions have fewer than 20 subjects.

introduce human biases [7]. To avoid these pitfalls, a common practice entails the wizard strictly adhering to a pre-defined interaction script.

The level of autonomy of the robot may also alleviate these issues: the more autonomous the robot, the smaller the human intervention surface, and the less likely the introduction of discrepancies between participants, given that a human operator will adapt their own behaviour in the interaction.

According to our findings (table III), around 40% of studies presented at the HRI conference over the last three years have implemented an interaction with a mostly autonomous robot. While this is certainly not negligible, it also means that a majority of the research presented at the HRI conference does not involve interactive autonomous systems.

To address this underlying misunderstanding caused by the differing high-level research goals, and in line with our goal to establish a common framework, one recommendation would consist of explicitly commenting in academic publications on the level of autonomy of the system, set in the perspective of the longer-term scientific agenda.

B. Participant Populations

For ecological validity it is good practice to perform evaluations with samples that are representative of the population with which a system is intended for use (i.e. to avoid *sample bias*). Such practice allows for better generalisation to the 'real-world', which is particularly desirable given that a large quantity of HRI research is conducted in the context of applications which require practicable solutions (autism therapy, child education, elderly care, etc.). There will undeniably be a trade-off between striving for ecological validity and experimental control, but there are a number of steps which can be taken with regards to participant populations that would be of great benefit to the validity of research in the field.

There is a clear imbalance of ages being used in HRI studies (figure 2). When research was not conducted with children (aged less than 18), or the elderly (aged over 65), 87% of studies used samples which drew from university populations (where age is stated). It may be the case that the intended end-user of these findings would indeed be only students/academic staff, or findings are not required to generalise to the wider population, but this seems unlikely to be the case for all instances. Additionally, it is worth noting that of the papers analysed that involved subjects, 18 did not report the age of these participants (figure 2), which further reduces the extent to which conclusions can be drawn.

Such samples are often dubbed 'convenience' samples, and whilst it is indeed convenient to use students which are readily available to test a system, questions must be raised as to how much can be gleaned from any findings. This will vary from case-to-case, but in principle, we feel that convenience samples should be avoided, as they may give rise to sample biases. We should strive towards greater ecological validity to push the field forwards, and ensure that the conclusions do not over-generalise away from the specific characteristics of the participant group used.

In addition, a substantial portion of evaluations in the field gather data from sample sizes which would be considered small in terms of human studies (figure 3). In psychology there have been concerns over small sample sizes leading to underpowered studies, in turn creating an incoherent body of literature [8].

For HRI to avoid these same problems, larger and more representative samples are required. However, this is not so easy to put into practice due to the sheer amount of effort involved in obtaining not just a greater number of participants, but also more diverse ones to maintain the generality of conclusions (where this is appropriate). Indeed, in some cases (e.g. in therapeutic or medical domains), larger sample sizes may not be possible. In this case, the importance of reporting standards come to the fore.

C. Evaluation Environments

The environment in which an evaluation is run can have a great influence on the behaviour and responses of participants [9]. The majority of studies in HRI appear to be run in laboratories, with an average of M=75% (SD=1%) of experiments conducted in the lab over the last three years of HRI conference publications. It has been debated within psychology as to whether lab experiments provide external validity (the extent to which generalisation to other settings and samples is possible) [10], with the conclusion that experiments at least require 'experimental realism': the degree of authenticity with regards to the phenomenon under exploration.

However, there is clearly a motivation for HRI experiments to move out of the lab and into the field, or the 'wild', in order to gather results which have demonstrable applicability. With such a commitment to field studies, there comes a trade-off between control and ecological validity. Some of these issues have previously been discussed in the context of HRI [9]. On the one hand, there is significant effort required on the part of the experimenters to run studies outside the lab, which needs to be acknowledged. Naturally however, the level of effort does not in itself guarantee a good study. Indeed, there is the possibility of introducing a number of new confounds related to the environment itself: for example the potentially complex effects of children talking to each other about the robot whilst the experiment is taking place in a school study.

As with the participants themselves (section IV-B), we suggest that ecological validity should be the main concern: is the experimental environment suitable given the experimental hypotheses? Secondly, we would suggest that since some types of confound are difficult to control for, a minimal requirement should be to report those confounds most likely to have an effect on the hypotheses.

D. Length of Empirical Studies

Novelty has often been raised as a potentially confounding or influencing factor for HRI studies [11], [12]. There is commonly a call for more long-term studies, or a statement of the desire for long-term investigation in the 'future work' section of HRI research papers. Table I shows that from 96 studies in the last 3 years, only 5 have consisted of more than one session interacting with a robot (one in 2014 and four in 2015). Whilst it is recognised that many longer-term studies may be published in different venues (be they journals or other conferences), these figures still raise questions about how we should consider the length of empirical studies.

There are of course many situations in which researchers may either wish to explicitly exploit a novelty effect, or a novelty effect is simply not relevant for the hypotheses in question. However, given a general desire to see HRI systems applicable to, and deployed in, the real world (e.g. as consumer systems), the issue of how human interactant behaviour will change over time as the novelty effect wears off remains an open question, whether this novelty effect applies at the level of the individual with expectations shaped by the anthropomorphic features of the robot (one person interacting repeatedly with a single robot system) or at the societal level (as social robots become commonplace in the public domain). For example, at the individual level, there are some suggestions that once the novelty effect is overcome, the robot behaviour will need to be more than just believable at a shallow level and beyond the role played by the robot embodiment, thus raising the necessity for deeper models of cognition and human behaviour [13].

What then constitutes long-term HRI? We would suggest that this is linked to the overcoming of the novelty effect, which in turn is related to the robot, its behaviour, and the interaction context, as elements influencing the extent to which novel behaviours are preferred over familiar ones [14]. This non-standard concept of the novelty factor may prove problematic in terms of comparing different studies. However, one way of addressing this could be to develop and use reliable behavioural metrics (based on gaze and linguistic behaviours for example) for the characterisation of familiarity.

E. The Approach to Statistics

Null-Hypothesis Significance Testing (NHST) is the de-facto standard for evaluating the importance of results. In this process, one checks the hypothesis that the data distribution (comprising sample size, mean and standard deviation for normally distributed data for example) obtained from an intervention condition does not differ from the distribution from a control condition (the null hypothesis): if this hypothesis can be rejected (i.e. a p-value less than or equal to some threshold, typically 0.05), the result may be considered 'significant'. On the face of it, this provides a useful means of characterising the 'success' (or not) of a method or intervention. This state of affairs is reflected in the HRI papers in our sample: $\sim$95% (90 out of 96 studies, see table I) of the papers employ NHST and report p-values to support the conclusions.

However, in recent years there has been increasing criticism of the importance conferred onto this means of statistical analysis in multiple fields of research[1], e.g. [15]. Indeed, the problematic nature of NHST has been acted upon by certain psychology journals, which have effectively banned the use of it to rest the main results of manuscripts on, e.g. [16]. This reflects three main concerns (and others): the arbitrary threshold for significance, replication sensitivity, and lack of effect size information.

Firstly, significance is typically held at a p-value of 0.05 or less (or 0.01 in the biological sciences). This is an arbitrary threshold (1 in 20 chance) that persists for historical continuity rather than theoretical or empirical merits. Determining the utility and/or importance (and this is often how significance is treated) of the result based on such an arbitrary threshold seems flawed from the perspective of the scientific method. Secondly, empirical results have suggested, and simulation studies have shown, that the p-value is highly volatile in experiment replications, with a variation in an initially significant p-value in the range [0.00008,0.44], 80% of the time [17]. p-values are thus unreliable in the face of replication. Thirdly, p-values do not incorporate any information about effect sizes: a highly statistically significant result from the perspective of NHST does not relate to the size of the observed experimental effect, and thus can not be used alone to assess the importance/impact of the result.

Descriptive statistics is sensibly recommended as the first stage of data analysis: we suggest that an increased emphasis on this should form part of standard reporting practice to circumvent some of the issues raised above. As an extension to this, we thus recommend that a minimal requirement for reporting mean-based data from multiple conditions should be

[1]Note that NHST is rigourous and mathematically valid, and thus not intrinsically problematic - the issue is rather the interpretation of the result, and the meaning derived from it in experimental contexts.

the provision by authors of Confidence Intervals (CI's) [17], [18], where the 95% CI is typically used[2]. Whereas p-values vary to a great extent, CI's have been shown to be more reliable, with an 83% chance that replication will give a mean within the CI of the original experiment [19]. CI's also inherently provide information about the effect size, thus providing an additional benefit over the reporting of p-values alone.

A further approach that could be brought to bear on this problem is statistical modelling. While this is on occasion seen to merely be an alternative means of performing a statistical analysis, we suggest that it should rather be seen as a change of perspective. Rather than forming just another statistical test of significance, the purpose is to gain an incrementally better view of the phenomena under investigation. In the Bayesian modelling perspective for example, there is an emphasis on the accumulation of data, of integrating new observations with existing knowledge. Previous results help to form *priors* for example, which shapes the way new data is viewed. In this perspective, the role of experimental methodology takes on a more central importance – it becomes the means by which data may be consistently integrated into ever more reliable priors. Our focus on guidelines to form a common methodological frame of reference thus feeds into these efforts.

F. Replicability

Replication (conducting the same experiment anew) and *reproduction* (re-running analyses on the original data to validate results) are instrumental in weaving a solid and trustworthy scientific fabric. Concerns have been voiced over the replicability of results in the sciences [20]. A recent large-scale replication of 100 psychology studies resulted in only 36% of studies having significant results, while originally 97 of the 100 studies reported significance ($p < 0.05$). A looser, subjective definition of replication found that only 39% of results could be deemed as successfully replicated [21]. While no published evidence exists on the replication of HRI studies, it is likely that replication will be of a similar level, due to the many methodological parallels between HRI studies and psychology studies.

A first obstacle is the lack of replicability: HRI studies are often challenging to replicate due to the nature of robotic hardware, the experimental setup, and the particular platform, environment and participants used. Access to specific robotic hardware is often restricted, especially if hardware is rare, expensive or difficult to access – e.g. androids or bespoke platforms. In addition, publications often do not have a detailed methods section facilitating replication, and software is, despite increased attention for open source initiatives, not widely shared in the HRI community.

On the other hand, increasing the reproducibility of our studies is likely less of a challenge. It mainly calls for sharing datasets and/or results and the means of analysing them (e.g. data processing scripts). Whenever the datasets can not be made anonymous, privacy concerns are likely to arise: those may be alleviated with agreed consent from the participants that "their data may be used for academic purposes" and through adequate sharing methods within the community. Note that we observe in recent years a clear trend toward ensuring datasets are available for papers to be considered for publication (case in point, taken from the author guidelines of PLOSOne: *"PLOS will not consider submissions from which the conclusions are based on proprietary data")*. We can only encourage the HRI community to actively embrace this practice.

A second obstacle however is the lack of incentive to replicate or reproduce studies. Academic reward systems and the current reviewing culture favour novelty over replication. This not only leads to a lack of validation of results and claims, but leads the field to chase the novel and exciting, rather than confirming or –perhaps even more importantly– refuting claims. As [21] eloquently points out, "Innovation points out paths that are possible; replication points out paths that are likely; progress relies on both".

A possible solution might be to create a new outlet for replication studies: if a journal or conference would welcome brief publications on successful or unsuccessful replications, this would demonstrate that replication is valued and would incentivise the consolidation of HRI insights.

V. Discussion

Our identification of six characteristics of HRI studies, supported by recent conference publication trends, and our subsequent exploration, have led to the proposal of six recommendations. These are both specific *researcher-level* recommendations that can be readily and practically applied to ongoing empirical work and the reporting of these, and also more general *field-level* recommendations that apply to the level of the field rather than individual researchers (see table IV for a summary).

A. Interdisciplinary Methods and Tools

Given the diversity of discipline-specific motivations and goals (section II), there are a number of sources that emphasise the importance of a common or shared mission if interdisciplinary efforts are to succeed, e.g. [22]. At the level of the field, we caution against specifying a mission statement that is too specific in terms of application or method. Such an effort would be likely to provide exclusions from the field, which we would suggest is (at least currently) unnecessary. From this perspective the current (brief) mission statement listed on the HRI community website provides a suitably general outline of the field: "*HRI is the multidisciplinary study of human-robot interaction*". At the level of individual research contributions however (e.g. a single study, series of experiments, or project), we believe such a statement to be a necessity for clarity of hypothesis, coherency, and appropriateness of the methods and metrics employed to investigate them.

However, with such a broad mission statement as used by the HRI community website, there need to be structures in place to ensure coherence in the field and to promote cross-disciplinary

[2]The use of 95% is a similarly arbitrary threshold as the 0.05 threshold for NHST p-values. However, CI's only provide a descriptive perspective, and not a metric of significance in themselves, thus avoiding the threshold problem.

TABLE IV
A SUMMARY OF THE RECOMMENDATIONS, WITH OPERATIONAL SUGGESTIONS AT RESEARCHER-LEVEL (ON THE LEFT) AND AT THE FIELD-LEVEL (ON THE RIGHT), WHERE APPROPRIATE.

R1: State the motivation, context and long-term goal of the research	
State the end-goal of the research (e.g. therapy, cognitive modelling, etc)	Provision and curation of collaborative, open tools to facilitate shared understanding and best-practice
R2: Clarify the level of robot autonomy	
The level of robot autonomy and/or 'wizarding' should be specifically and clearly stated; wizarded robot behaviours should be avoided as a benchmark condition.	
R3: Use of ecologically valid subject groups and experiment environments	
Based on the experimental hypotheses, assess the appropriate subject group; recognise the constraints that the use of a single subject group imposes on the study conclusions	
R4: Relate the notion of long-term interactions to overcoming the novelty effect	
Introduce metrics for familiarity of the study subjects with the robot as a means of characterising the novelty effect	
R5: Use descriptive statistics	
As a minimal requirement, report 95% Confidence Interval for metrics of each condition; emphasise the build up of evidence over arbitrary significance judgements.	Enforce reporting standards in conference and journal publications
R6: Support replication and reproduction	
Ensure detailed methodology; provide source code whenever possible; publish datasets and/or intermediary results, along with the tooling to analyse them (when applicable)	Provision of a peer-reviewed publication venue specifically for independent experimental replications; provide guidelines and infrastructure to share datasets

collaboration while preventing fragmentation. We suggest above (section II) that the imposition of common benchmark tasks could introduce unwanted biases in the long-term, and introduce technical barriers to entry for certain sections of the community. Our proposal to formulate a common framework for methodological and reporting considerations forms the beginning of an alternative approach. In the same way that a characterisation methodology such as conversational analysis can provide a common and formal basis for comparison of qualitative observations between studies, so can such a common framework do the same for the multiple disciplines within HRI. The recommendations we propose (summarised in table IV) are pitched at two levels to encourage a coordinated effort at achieving this: standards for individual researchers to follow, but also suggested changes in field-level infrastructure that can bring about the wider cultural change desired to facilitate the efforts of individuals. Indeed, such efforts are apparent in other fields, for example in health research (equator-network.org).

Regarding this field-level infrastructure, the provision of a number of tools for collaboration and shared understanding would be of use in addressing some of the issues that arise from a vibrantly interdisciplinary field. One such tool is a community FAQ. Such a resource could contain technical advice/resources, reporting recommendations, explanations of key jargon, best practices, etc. covering all HRI disciplines (quantitative and qualitative, technical and social). This would contribute to bridging cross-cultural "language" issues by having one entry-point that researchers (and newcomers to HRI research in particular) could use as a reference.

However, as with any introduction of new standards and/or recommendations, there is a need to minimise the 'barrier to entry' to maximise uptake within the community. The more specific researcher-level recommendations we make are pitched to minimise this barrier, whilst providing significant benefits. Our recommendations for collaborative tools and field-level infrastructure (publication support for peer-reviewed replication studies for example) on the other hand will require more significant personal investment, although if such tools are mandated as part of article submission processes (for example), the motivation to conform is likely to prove sufficient to overcome any initial inertia.

B. Facilitating Long-Term HRI

One feature raised from recent studies is a notably small number of longer-term studies (section IV-D). Since novelty effects are typically present in shorter-term evaluations, and given the as yet under-appreciated role that robot morphology design plays in shaping interaction expectations, it is difficult to assess from current evidence what long-term phenomena arise in genuinely long-term interactions between humans and robots. In this case, there is a strong drive to increase the autonomous competencies of the robots that are able to support these studies. However, our paper review exercise has shown, commensurate with the interdisciplinary nature of the field of HRI, that levels of autonomy in robotic systems are currently only limited (section IV-A). This clearly represents a significant challenge for the community: with the requirement for autonomous behaviour comes a need for more elaborated models of appropriate robot behaviour generation in response to social and environmental cues. Efforts in this area are becoming increasingly prevalent in the fields of AI and Cognitive Science, with a multitude of cognitive architectures being developed [23], although these have as yet only a limited impact in HRI.

This requirement for deeper levels of cognitive model is not in our view restricted to the more robot-centred strands of HRI; we suggest it is also a central requirement for the human-centred perspectives. There is a need to formalise in some way the knowledge of human behaviour and adaptation (including psychology, cultural studies, and neuroscience to

varying degrees) to enable application to HRI, whether it is in the form of a robotic system, or as a means of analysing human behaviour (whether it be reaction times or learning outcomes) in an experimental setting.

C. Discipline Dependencies

From the outset of this paper, we have emphasised that HRI lies at the convergence of multiple disciplines; we have also suggested that it would be beneficial to maintain this plurality of approaches. However, we must then also acknowledge that these different disciplines have differing dependencies and goals (section II).

For example, technical developments have the power to advance the field. Given the central role of robots in HRI (in all senses of the phrase), this is uncontroversial. However, there are mutual constraints on these developments. For example, as we have shown (section IV-A), robot wizarding is partially employed to overcome various technical challenges, which results in a limited capacity to engage in long-term studies (section IV-D). Whereas technically-oriented papers may typically appear in other publication venues, the more recent introduction of the technical theme in the HRI conference reflects an acknowledgement of this dependency on technical issues. Nevertheless, it may be worth raising the expectations of the technical content of all HRI contributions as part of the review process, in the same way that methodological issues are currently rigorously assessed.

There of course remain further open questions in the field that will require multi-disciplinary consideration. One notable example of this is the role that robot behaviour and morphology relate to one another with respect to human perceptions and reactions. Such theoretical and design questions are clearly fundamental to overall progress in the field, including to applications. We suggest that the resolution to these issues, and others, will require the application of empirical investigation to characterise and explore the phenomena: i.e. conducting studies to collect data to subsequently inform further refinement. Our focus in this paper on providing a common frame of reference through methodological guidelines is precisely aimed at providing support for such multi- and cross-disciplinary efforts: our recommendations (table IV) provide the basis of this frame of reference.

VI. Conclusion

What we advocate for the field of HRI is the maintenance of the plurality of discipline-specific motivations, rather than the imposition of a single set. Nevertheless, a common framework should be provided to facilitate the interaction of these differing approaches such that the non-unitary field as a whole can move forward. In other words: to maintain HRI as a collaborative field between disciplines, rather than their unification into a new single field. In this paper, we have examined recent trends in HRI publications to define challenges that face this interdisciplinary approach, and derived both practical and more general methodological recommendations that we suggest will provide the start of a much needed common frame of reference that will consolidate the progress made thus far, and provide a platform for future contributions.

Acknowledgements

This work is partially funded by the EU FP7 project DREAM (grant 611391, http://dream2020.eu), and the EU H2020 project L2TOR (grant number 688014, http://www.l2tor.eu).

References

[1] J. L. McClelland, "The Place of Modeling in Cognitive Science," *Topics in Cognitive Science*, vol. 1, no. 1, pp. 11–38, Jan. 2009.

[2] A. Tapus, M. J. Mataric, and B. Scassellati, "The Grand Challenges in Socially Assistive Robotics," *IEEE Robotics and Automation Magazine*, vol. 14, no. 1, pp. 35–42, 2007.

[3] B. D. Argall and A. G. Billard, "A survey of Tactile Human-Robot Interactions," *Robotics and Autonomous Systems*, vol. 58, no. 10, pp. 1159–1176, 2010.

[4] Y. Nagai and K. J. Rohlfing, "Computational analysis of motionese toward scaffolding robot action learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 44–54, 2009.

[5] C. Bartneck, "The end of the beginning: a reflection on the first five years of the HRI conference," *Scientometrics*, vol. 86, no. 2, pp. 487–504, 2011.

[6] L. D. Riek, "Wizard of Oz studies in HRI: A systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, 2012.

[7] I. Howley, T. Kanda, K. Hayashi, and C. Rosé, "Effects of social presence and social role on help-seeking and learning," in *9th ACM/IEEE International Conference on Human-Robot Interaction (HRI'14)*. ACM, 2014, pp. 415–422.

[8] S. E. Maxwell, "The persistence of underpowered studies in psychological research: Causes, consequences, and remedies," *Psychological Methods*, vol. 9, no. 2, p. 147, 2004.

[9] R. Ros *et al.*, "Child-robot interaction in the wild: Advice to the aspiring experimenter," in *13th International Conference on Multimodal Interfaces (ICMI'11)*. ACM, 2011, pp. 335–342.

[10] L. Berkowitz and E. Donnerstein, "External validity is more than skin deep: Some answers to criticisms of laboratory experiments," *American Psychologist*, vol. 37, no. 3, p. 245, 1982.

[11] R. Gockley *et al.*, "Designing Robots for Long-Term Social Interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)*. IEEE, 2005, pp. 1338–1343.

[12] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial," *Human-Computer Interaction*, vol. 19, no. 1, pp. 61–84, 2004.

[13] S. Lemaignan, J. Fink, P. Dillenbourg, and C. Braboszcz, "The Cognitive Correlates of Anthropomorphism," in *Proceedings of the Workshop: A bridge between robotics and neuroscience at the Human-Robot Interaction Conference*, Bielefeld, Germany, 2014.

[14] I. Leite, C. Martinho, and A. Paiva, "Social Robots for Long-Term Interaction: A Survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[15] R. Nuzzo, "Statistical errors," *Nature*, vol. 506, no. 7487, pp. 150–152, 2014.

[16] D. Trafimow and M. Marks, "Editorial," *Basic and Applied Social Psychology*, vol. 37, no. 1, pp. 1–2, 2015.

[17] G. Cumming, "Replication and p Intervals: p Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better," *Perspectives on Psychological Science*, vol. 3, no. 4, pp. 286–300, Jul. 2008.

[18] D. H. Johnson, "The Insignificance of Statistical Significance Testing," *The Journal of Wildlife Management*, pp. 763–772, 1999.

[19] G. Cumming, J. Williams, and F. Fidler, "Replication and researchers' understanding of confidence intervals and standard error bars," *Understanding Statistics*, vol. 3, no. 4, pp. 299–311, 2004.

[20] R. D. Peng, "Reproducible Research in Computational Science," *Science*, vol. 334, no. 6060, pp. 1226–1227, 2011.

[21] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, p. aac4716, 2015.

[22] R. R. Brown, A. Deletic, and T. H. Wong, "How to Catalyse Collaboration," *Nature*, vol. 525, pp. 315–317, 2015.

[23] D. Vernon, *Cognitive Systems - A Primer*. MIT Press, 2014.

Heart vs Hard Drive: Children Learn More From a Human Tutor Than a Social Robot

James Kennedy, Paul Baxter, Emmanuel Senft and Tony Belpaeme
Centre for Robotics and Neural Systems
Plymouth University, U.K.
{james.kennedy, paul.baxter, emmanuel.senft, tony.belpaeme}@plymouth.ac.uk

***Abstract*—The field of Human-Robot Interaction (HRI) is increasingly exploring the use of social robots for educating children. Commonly, non-academic audiences will ask how robots compare to humans in terms of learning outcomes. This question is also interesting for social roboticists as humans are often assumed to be an upper benchmark for social behaviour, which influences learning. This paper presents a study in which learning gains of children are compared when taught the same mathematics material by a robot tutor and a non-expert human tutor. Significant learning occurs in both conditions, but the children improve more with the human tutor. This difference is not statistically significant, but the effect sizes fall in line with findings from other literature showing that humans outperform technology for tutoring. We discuss these findings in the context of applying social robots in child education.**

I. Introduction

An increasing quantity of research in HRI has considered the use of robot tutors, particularly for educating children [1], [2], [3]. It has been found that robot embodiment [1], [2], social behaviour [4], and teaching strategies [3], [5] can improve child learning. One question that often arises, particularly from non-academic audiences, is how robots compare to human tutors. The aim of such research is rarely to replace human teaching, but to supplement it, so such a comparison is not typically part of experimental hypotheses.

Given the link between robot social behaviour and learning [4], human behaviour is often used to derive behaviour for robots to provide an upper benchmark of social behaviour that robots can aim for in tutoring. The literature from other fields suggests that human tutoring also provides an upper benchmark in terms of learning gains [6], but this has not been verified in HRI. Serholt *et al.* [7] found no significant difference between the performance of children who had been tutored by a humanoid robot compared to a human, but the robot speech was controlled using a Wizard-of-Oz method, introducing additional variability between conditions. The present paper reports on a study in which the lesson content delivered by a human and an autonomous robot is kept consistent in order to explore the differences in child learning depending on the character (and their social behaviour) providing the content. The aim is to address the following hypothesis:

H1: Human tutoring will lead to more child learning when compared to robot tutoring.

II. Methodology

The study employs the same methodology as seen in [2] and [4]. Children aged 8 and 9 engage in a dyadic interaction in their school with a tutor who guides them through a method for prime number identification. The children's learning is measured through a pre-test and a post-test consisting of 12 numbers which need to be categorised as 'prime' or 'not prime' (6 per category). Prior to the interaction, children have not learnt about prime numbers, but the technique relies on their ability to divide by 2, 3, 5 and 7, so this is also tested. The tutor provides hints to help with the division, as well as a lesson about how to identify prime numbers using the Sieve of Eratosthenes technique. Two tests for prime number identification are used in a cross-testing strategy to control for exposure to the tests.

Two conditions were employed: *(1)* an autonomous 'high immediacy' robot tutor [4], and *(2)* a human tutor (Fig. 1). The robot tutor was designed to regularly gesture, look at the child, make small body movements to appear 'relaxed', and lean forwards. The human was given a word-by-word script to match the lesson content of the robot, but was not constrained in terms of social behaviour. Due to the script providing precise lesson content (and the study focus on social behaviour and embodiment differences) an expert tutor was not required. A total of 22 children took part: 11 in the robot condition and 11 in the human (age M=8.8, SD=0.4; 12F, 10M). Interactions lasted for M=14m05 (SD=3m16) in the robot condition, and M=13m10 (SD=3m39) in the human condition.

III. Results and Discussion

Children improve significantly in both conditions (Fig. 2). Paired t-tests show the post-test score (M=7.6, 95% CI [5.5,9.8]) is significantly higher than the pre-test score (M=5.2, 95% CI [3.7,6.7]) in the human condition; $t(10)$=2.425, p=.036. The post-test score (M=7.0, 95% CI [4.9,9.1]) is also significantly higher than the pre-test score (M=5.1, 95% CI [3.4,6.8]) in the robot condition; $t(10)$=3.057, p=.012. Although the children improve more between the pre-test and post-test in the human condition (M=2.5, 95% CI [0.2,4.7]) than in the robot condition (M=1.9, 95% CI [0.5,3.3]), this difference is not found to be statistically significant using an independent samples t-test; $t(20)$=0.459, p=.652.

The improvement from pre- to post-test score is not significantly different between the robot and human conditions, but

978-1-4673-8370-7/16/$31.00 © 2016 IEEE

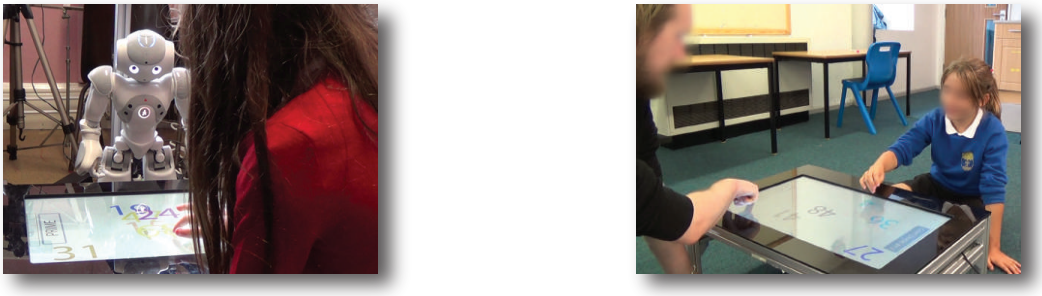


Fig. 1. Images of the interactions: (*left*) the robot condition, (*right*) the human condition. Interactions take place around a touchscreen which displays the learning material. Both the child and the tutor (whether human or robot) can move numbers on the screen. Feedback is provided by the tutor, and not on screen.

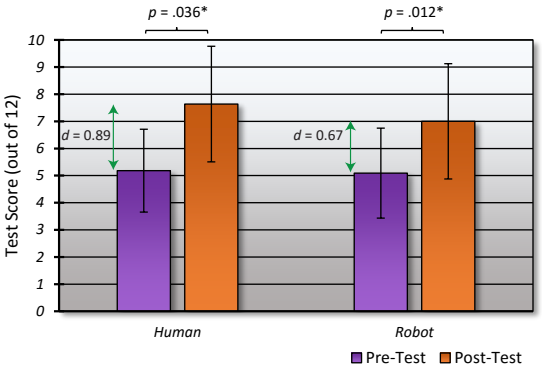


Fig. 2. Child pre-test and post-test scores for the robot and human conditions. The improvement is significant in both conditions, showing that the children learn. The difference between conditions is not significant, but the improvement effect size is larger in the human condition. *Error bars* show the 95% Confidence Interval.

this may be due to the relatively small sample size (although *t*-test assumptions are met). If the trends here were to continue, then this difference would become significant with more subjects. The effect size seen in each condition provides a clearer indication of the difference between them; Cohen's d=0.67 for the robot, but d=0.89 for the human. As such, this provides some support for H1: that child learning is greater when tutored by a human when compared to a robot. These effect sizes are similar to those found in other literature [6]. It should be noted that the effect sizes in [6] compare to a no tutoring control, which is not done here since the nature of the task makes learning unlikely without tutoring.

It is also worth noting that the human condition mean was lowered by one instance where the child had clearly learnt the technique, but confused the categories, and so scored 0 on the post-test (i.e. 100%, but incorrect). The child asked for clarification, but as this help would not have been available in the robot condition, it was not given by the human at the time.

The specific robot and human used in the tutoring task will have had a large impact on the results. One robot (and its behaviour) was compared to one human; these results are likely to vary depending on the robot and human used. The learning content was kept consistent between the conditions, but the social behaviour was not constrained in the case of the human. This means that the human can take advantage of some social cues that the robot could not, and could subsequently be more socially adaptive (for example, in mutual gaze) than the robot, which may account for some of the learning differences. A non-expert human was used due to the tightly specified learning content, but an expert tutor may have used different social behaviour, potentially leading to more learning. It remains to be seen if the robot could close the gap in learning outcomes with improved social sensitivity and behaviour.

Of course, the aim is not to replace human tutors; robots offer additional opportunities to supplement current human tutoring provision. Robots can assume a wider variety of roles, for example, to assist teachers [1], or to offer children a chance to teach a less-able peer [3], [5]. Alternatively, robots could provide personalised support which falls outside of typical lessons or the school environment, such as additional language support for non-native children, as discussed in [8].

IV. Acknowledgements

This work is partially funded by the EU FP7 DREAM project (grant 611391), the H2020 L2TOR project (grant 688014), and the SoCEM, Plymouth University, U.K.

References

[1] M. Alemi *et al.*, "Employing Humanoid Robots for Teaching English Language in Iranian Junior High-Schools," *Int. Journal of Humanoid Robotics*, vol. 11, no. 3, 2014.

[2] J. Kennedy *et al.*, "The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning," in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI*. ACM, 2015, pp. 67–74.

[3] F. Tanaka and S. Matsuzoe, "Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 78–95, 2012.

[4] J. Kennedy *et al.*, "Higher Nonverbal Immediacy Leads to Greater Learning Gains in Child-Robot Tutoring Interactions," in *Proc. of the Int. Conf. on Social Robotics*. Springer, 2015, pp. 327–336.

[5] D. Hood *et al.*, "When Children Teach a Robot to Write: An Autonomous Teachable Humanoid Which Uses Simulated Handwriting," in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI*. ACM, 2015, pp. 83–90.

[6] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," *Educational Psychologist*, vol. 46, no. 4, pp. 197–221, 2011.

[7] S. Serholt *et al.*, "Comparing a humanoid tutor to a human tutor delivering an instructional task to children," in *Proc. of the 14th IEEE-RAS Int. Conf. on Humanoid Robots*. IEEE, 2014, pp. 1134–1141.

[8] T. Belpaeme *et al.*, "L2TOR - Second Language Learning Tutoring using Social Robots," in *Proc. of the First Int. Workshop on Educational Robots at ICSR'15*, 2015.

Social Robot Tutoring for Child Second Language Learning

James Kennedy, Paul Baxter, Emmanuel Senft and Tony Belpaeme
Centre for Robotics and Neural Systems
Plymouth University, Plymouth, U.K.
{james.kennedy, paul.baxter, emmanuel.senft, tony.belpaeme}@plymouth.ac.uk

***Abstract*—An increasing amount of research is being conducted to determine how a robot tutor should behave socially in educational interactions with children. Both human-human and human-robot interaction literature predicts an increase in learning with increased social availability of a tutor, where social availability has verbal and nonverbal components. Prior work has shown that greater availability in the nonverbal behaviour of a robot tutor has a positive impact on child learning. This paper presents a study with 67 children to explore how social aspects of a tutor robot's speech influences their perception of the robot and their language learning in an interaction. Children perceive the difference in social behaviour between 'low' and 'high' verbal availability conditions, and improve significantly between a pre- and a post-test in both conditions. A longer-term retention test taken the following week showed that the children had retained almost all of the information they had learnt. However, learning was not affected by which of the robot behaviours they had been exposed to. It is suggested that in this short-term interaction context, additional effort in developing social aspects of a robot's verbal behaviour may not return the desired positive impact on learning gains.**

***Index Terms*—Human-robot interaction; robot tutors; second language learning; social availability; immediacy**

I. Introduction

An increasing number of human-robot interaction (HRI) researchers are exploring the utility of robots for tutoring children [1], [2], [3]. Much of this research is centred around the social behaviour of the robot, with a view to improving learning outcomes and child responses to the robot [4], [5]. However, there are still many questions to be answered about how a robot should behave in educational interactions in order to achieve these goals [6].

Social interaction has been highlighted as a particularly important element in language learning [7], and recent research in HRI suggests that robots are able to make a positive impact on learning in such contexts [1], [8]. One aspect of social interaction which is positively correlated with learning between humans is the 'psychological availability' of an instructor [9], [10], [11]. Certain elements of 'availability' in social behaviour have been studied in HRI before [12], [13], but an explicit effort to manipulate this availability and examine the effect on child learning remains to be carried out.

Child language learning provides an ideal domain for social HRI to contribute to. In the case of language, children learn better than adults, despite the increased cognitive capacity of adults. Language learning has a 'critical period' in neurobiology [14], which means that there is a window in which it is best learned. As such, in this paper we conduct a study with children aged 8 and 9 years old. At this age, the children are still within the critical period, but have sufficient skill to read novel words without assistance.

We aim to explore how the language learning of a child can be influenced by the social behaviour of a robot tutor. This paper presents an experiment in which a robot tutor teaches children some aspects of a second language. The robot behaviour is modified to be more or less socially available through the verbal interaction it has with the child. The learning of the children is measured in the short-term (immediately after the interaction), and also the following week to check that the learned information was retained. We seek to investigate whether the intended availability of the robot is perceived by the children, and whether a more socially available robot has a positive impact on learning outcomes as predicted by the HRI and human-human interaction (HHI) literature.

II. Related Work

A. Language Learning with Robots

Social robots have proven their utility in language learning environments with improved outcomes when teaching is supplemented with robots [1], [2]. Alemi *et al.* [1] used a NAO robot in a school classroom to support a human teacher in teaching English as a foreign language. Knowledge was assessed before and after 5 lessons (one per week for 5 weeks). It was found that children in the condition with a robot learned and retained significantly more vocabulary than children who had a human teacher alone.

However, things are not as clear when the robot is interacting one-on-one with students without a human teacher present. Various experiments have sought to apply human-human learning principles to child-robot interactions in the language domain with mixed results [8], [15]. Curiosity of a robot was used to inspire reciprocal behaviour in children as the HHI literature predicts an increase in learning when children are more curious. Although the children who saw the curious robot adopted curious behaviours, their word learning did not improve any more than those children who had not seen the curious robot [4].

Some effects have been successful though: a robot with personalised story-telling complexity resulted in children using more words and more diverse words than children

who interacted with a non-personalised robot [15]. Socially supportive behaviours have also successfully been implemented in a robot which taught a novel language to children [16]. Those in the socially supportive condition scored significantly higher on a language test and in motivation measures (intrinsic and task motivation). The socially supportive condition employed many non-verbal behaviour manipulations, such as increased empathy, attention guiding, and non-verbal feedback. Whilst this is a promising result, more needs to be done to establish solid models for robot social behaviour in interactions of this nature. This paper seeks to address how the verbal social behaviour of a tutor robot affects child learning and how such behaviour might be characterised.

B. Social Behaviour and Learning

In order to maximise the potential of robots in learning contexts, it is useful to explore how they should behave socially, as many human-human studies have revealed a link between social behaviour and learning [10], [11], [14]. Social behaviour also has a great impact on how students perceive teachers [10], [17]. In turn, this influences factors such as how much students believe they have learnt, and how motivated they are to learn [11]. Therefore, it is important for students interacting with robots in educational contexts to have a positive perception of, and relationship with, the robot.

One concept of human social behaviour which has been positively correlated with student motivation, student achievement, and student attitudes is the 'psychological availability' of an instructor [10], [11]. This concept considers how a teacher acts towards any particular pupil (as opposed to the class as a whole, given the classroom context of many studies in this field). This availability is measured through 'immediacy' and consists of verbal and nonverbal social behaviour components [9], [18]. It should be noted that typical connotations of the word 'immediate' regarding timing do not form part of the measure. Instead, verbal immediacy includes behaviours such as whether an instructor uses personal examples in teaching, uses first names, solicits student opinions, and so on, whereas nonverbal immediacy considers the use of overt nonverbal social cues such as gaze and gestures [9], [17].

Research has been done in HRI with a view to improving the bond between children and robots through some of these means [19], although often not in the context of educational interactions. It has been found that 'off-activity talk' - dialogue with a robot which does not concern the task being completed - encourages compliance in children in a therapeutic setting [13]. Personalisation in therapeutic contexts has also been considered. Children were asked a number of questions about their preferences and the robot then mentioned these in an interaction, the children who interacted with a personalised robot enjoyed the interaction more, but subject numbers were too low for statistical comparisons [12].

Part of the social availability construct (nonverbal immediacy) has previously been used in HRI with findings in agreement with the HHI literature [20], [21], suggesting immediacy is suitable for use as a metric in HRI. This paper considers the other part of the social availability construct, verbal immediacy, to measure and motivate robot behaviour differences.

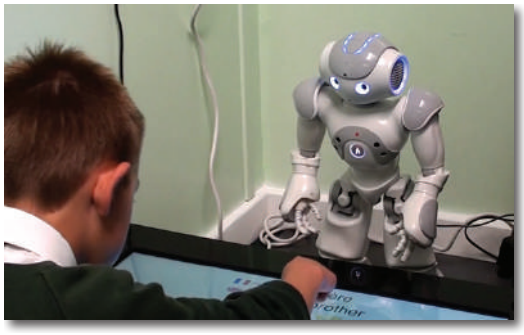


Fig. 1. A child answering a question on screen during the interaction.

III. Research Questions

Following on from previous research with humans [9] and robots [12], [13] we seek to test whether robot verbal availability has a positive impact on children's second language learning as predicted by the literature. In order to make such an assessment, it first needs to be clear that children perceive the behaviour of the robot as intended. Verbal immediacy provides a basis for measuring the children's perceptions and also for motivating differences between robot conditions. To ensure that any observed learning effects are retained and not just the product of short-term memory recall, we also aim to verify children's retention of the material outside of the short-term interaction context (as in [22]). This leads to the following hypotheses:

- H1. *Perception of robot behaviour.* Children will perceive and report differences in the robot's verbal availability (as measured through immediacy).
- H2. *Child learning.* Children will retain the language skills that they learn from the robot outside of the short-term.
- H3. *Effect of availability on learning.* A robot exhibiting more socially available verbal behaviour will lead to greater child learning gains than a robot without this behaviour.

IV. Design

French is commonly taught in English schools, so would have clear relevance for the children. However, it does not receive very much lesson time (the majority of schools offer 30-45 minutes per week at the age used in this study [23]), so there is plenty of scope to teach new concepts. As such, French was selected as the second language to teach in this study. The learning material was developed in collaboration with an academic researcher in language development, a native French speaker, and a teacher.

The structure of the lesson content was designed based on previous work in which children learnt mathematical concepts, such as [5], and a pilot study involving a human tutor and children. The aim was for the children to learn that nouns in French have a gender, that this changes which article is used

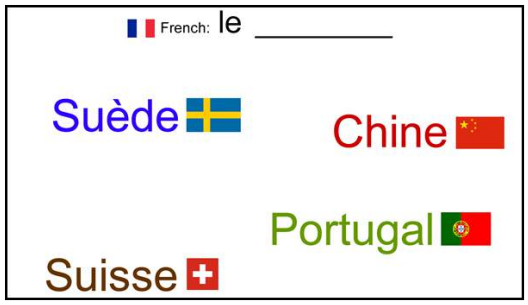


Fig. 2. Screenshot from the touchscreen showing a question. Children can touch a word, drag it to the blank space and release to answer. Here the correct answer being 'Portugal'.

('le' or 'la'), and that for some words there are patterns which can be used to help work out which article to use.

An Aldebaran NAO robot acted as a tutor, delivering all lessons through speech and moving words on a touchscreen (Fig. 1). As such, the children were exposed to both the words' pronunciation and orthography. The robot demonstrated how questions could be answered by dragging and dropping the correct answer in the blank space (see Fig. 2). The robot first explained the concept of words having a gender by using an English example (using 'waiter' for a man, and 'waitress' for a woman). Following this, it explained how the French word for 'the' could be 'le' or 'la' depending on the gender of the noun it precedes. The robot then explained rules for working out whether to use 'le' or 'la'. After explaining each rule, the child's understanding was checked (Fig. 3).

During the lessons the robot would explain a rule and then use the screen to show an example. The rules used were taken from online French language learning guides[1,2] and were verified by a French native speaker. The rules were as follows: 1) 'le' is used for male people, and 'la' is used for female people, 2) 'la' is used for countries ending in 'e', 3) 'la' is used for fruit or vegetables ending in 'e'. Whilst these are recognised techniques for people learning a second language, it should be made clear that it is unlikely that a native speaker would learn in this way, and that there are a limited number of exceptions to rules 2 and 3 (but these were avoided in the lesson content here). We do not seek to determine the best teaching strategy for the concept, but the effect that robot behaviour has on any learning.

Questions were designed to get progressively more complex as the interaction progressed. To start with, English translations and pictorial representations of the words were provided alongside the French. At this stage, the child was only required to select the article 'le' or 'la' to add to the word. Towards the end of the interaction, all English translations were removed so that only the French and the pictures remained. The question structure was also changed in later stages: the child was required to match a noun to the article (Fig. 2), which requires them to

[1]http://goo.gl/JPjmPO
[2]https://goo.gl/WY37z5

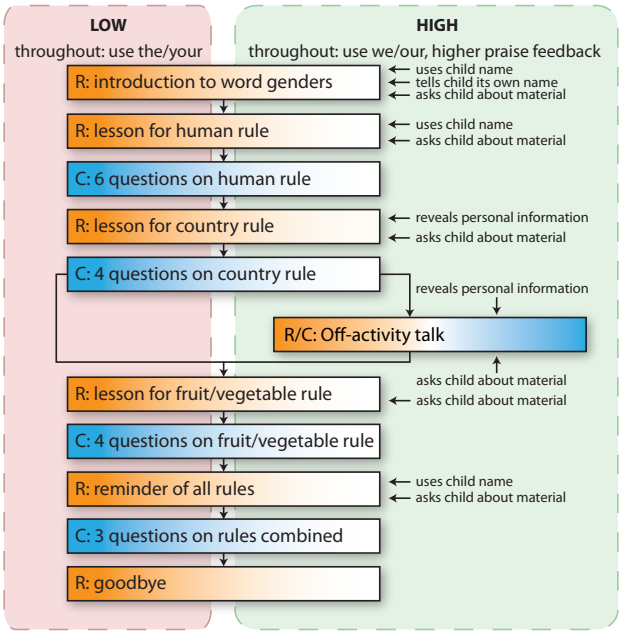


Fig. 3. Structure of the task. *R* refers to robot explanation sections and *C* refers to child question answering sections. The robot dictates the structure of the interaction through speech and by presenting questions on the touchscreen, informing the child of when it is their turn answer questions on the screen. The HIGH condition includes many manipulations in the verbal behaviour to make it more 'available'.

assess several nouns for each question, rather than just one as in the earlier questions.

All feedback was provided verbally by the robot; no feedback was shown on the screen. When providing feedback, the robot's TTS would switch to French so that the child could hear the correct pronunciation. The robot was autonomous throughout, except for some short vocal phrases in one condition, which were triggered by the experimenter (see Section V-C).

V. Evaluation

A. Participants

A total of 67 children were included in the study after exclusions due to technical issues (1 child) or absence from school during one of the two visiting periods (7 children). All children were native English speakers and were from the same year group (with three class teachers) from a primary school in the U.K. (average age M=8.8, SD=0.4; 30M, 37F). Only one child was fluent in another language (this language was not used in this study). Children were distributed randomly between groups whilst balancing for gender and class teacher. All children had parental/guardian permission and gave their consent to take part in the study.

B. Measures

Learning was measured through pre-, post- and retention tests, which can be seen online[3]. These tests sought to examine various aspects of the children's learning, including their

[3]http://goo.gl/hrIQEe

vocabulary acquisition, and their ability to apply each of the 3 rules in isolation and combination with each other. The test consisted of 12 questions: 3 vocabulary-based (1 French-English and 2 English-French), 2 about humans (rule 1), 2 about countries (rule 2), 3 about fruits and vegetables (rule 3), and 2 combined all three rules. Each question had 4 multiple choice answers and used the same formats as questions on the touchscreen. The majority of the test questions used words that the children had not seen in the learning material in order to ensure generalised learning was taking place, rather than memorisation of specific instances; exceptions are discussed in Section VII. The pre-, post- and retention-tests were all the same as this was necessary to account for children's prior knowledge (they had learnt some French vocabulary in school before), and to accurately measure their recall. The children were not given any feedback on their tests at any stage.

The child's perception of the robot was measured through a questionnaire combining verbal immediacy and nonverbal immediacy items. This 23 question questionnaire was completed on paper and was multiple choice. The verbal immediacy and nonverbal immediacy items were based on those used in [10], but were modified such that the language could be understood by children. The final questionnaire used can be seen online[4]. Verbal immediacy includes aspects of behaviour such as personalisation, off-activity talk, and student opinion solicitation. Nonverbal immediacy covers overt social behaviours, such as whether gestures are used, whether the robot looks at the child, and so on.

C. Conditions and Robot Behaviour

In order to address the hypotheses in Section III, three conditions were devised: 1) a robot with high verbal availability (HIGH, n=20), 2) a robot with low verbal availability (LOW, n=20), 3) a control with no robot and just a pre- and retention test (CTRL, n=27). The robot with low verbal availability doesn't have the verbal behaviours which lead to being considered available as measured by verbal immediacy (Fig. 3). The control condition is used to verify that there are no learning effects due to exposure to the test material.

In both robot conditions, the nonverbal behaviour was kept constant. The behaviour used was designed to be of high nonverbal immediacy, with the robot's gaze randomly moving in the direction of the child, gestures during speech, a slight lean forward of the body, and slight motor noise in the arms to give the impression of being relaxed. The perception of this behaviour as being of high nonverbal immediacy is verified through the questionnaire after the interaction (as described in Section V-B).

The speech of the robot was kept the same in both conditions outside of the experimental manipulations as described below. This ensures that the lesson content is largely unchanged between conditions, although the experimental manipulations require some language adjustments, these should not impact on the coherence or intelligibility of the lessons.

Verbal immediacy can be used to measure aspects of availability of an instructor, so the verbal immediacy questionnaire [9] was used to create the robot conditions with different availability levels. In order to generate the behaviour for the conditions, we therefore applied all of the verbal immediacy questionnaire items possible to the speech for the HIGH condition, and did not apply them for the LOW condition. The following differences were present in the HIGH condition robot behaviour, but not in the LOW condition[5]:

1) use the child's name (3 times)
2) tell the child its name
3) reveal personal information about itself (twice in addition to its name)
4) ask the child how they felt about the material (e.g. "does everything make sense to you so far?" 6 times)
5) ask the child about their hobbies and continue the discussion for 2 or 3 speech turns
6) use "we/our" work (as opposed to "the/your", throughout)
7) provide higher praise feedback (e.g. "You're doing really well! That was right", as opposed to simply "That was right" in the LOW condition)

Two items of the verbal immediacy questionnaire were not manipulated: humour and feedback provision. Humour was considered to be inappropriate to add given the context of the interaction and difficulties in selecting a comment that would be universally funny. Whether or not feedback was provided was not manipulated between conditions as in this context, the only way of getting feedback was from the robot and missing feedback here would confound any findings related to learning.

To compensate for unreliable speech recognition, a Wizard-of-Oz intervention was used in the HIGH condition to let the robot reply 'that's great' after the children answered a question from the robot about their understanding of the material (children always said they had understood the lesson), and to trigger pre-scripted phrases at the appropriate time for the discussion about the child's hobby.

D. Procedure

The interactions took place on the school premises in a quiet working space familiar to the children. The child sat across from an Aldebaran NAO with a 27 inch touchscreen placed horizontally between them (Fig. 1). Two video cameras were used to record the interactions. One experimenter sat behind and to the side of the child, out of their view (Fig. 4). The time children spent interacting with the robot was on average M=11min 26s (SD=1min 11s).

The experimenter spent a full week in the school, plus one day the following week. On the first Monday of the visit, pre-tests were delivered to all children in their main classrooms. These were completed under the supervision of the experimenter and the class teacher to make sure that children completed them individually. Throughout the week those children interacting with the robot would be taken out of class individually, take part in the interaction, and then

[4]http://goo.gl/UoL5QM

[5]Please also refer to the video figure for this publication

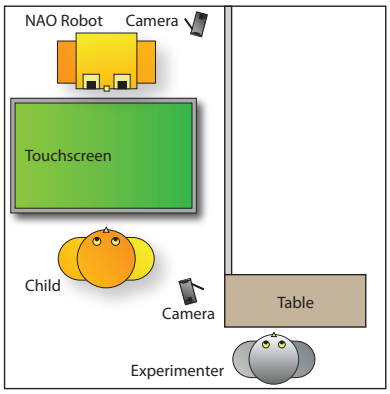


Fig. 4. Schematic overview of the interactions being investigated in this paper. The child and the Aldebaran NAO robot sit across a touchscreen from one another. An experimenter sits behind and out of view of the child. Two video cameras record the interaction. Figure not to scale.

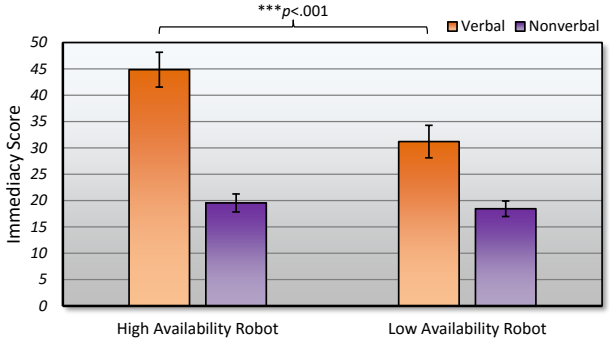


Fig. 5. Verbal and nonverbal immediacy scores for the high availability (HIGH) and low availability robot (LOW) conditions. The HIGH condition is perceived to have significantly higher verbal immediacy while having the same nonverbal immediacy, showing that the children perceive it as more available. *Error bars* show 95% CI.

complete the post-interaction test and questionnaire on paper, to the side of where the experimenter had been sitting (so they can no longer see the robot or touchscreen). The robot condition was switched between each interaction to ensure a balance throughout the week.

On the Monday of the following week the experimenter returned to deliver the retention test to the children under the same conditions as the pre-test. Children in the control group therefore completed a pre-test and a retention test without any teaching input. The children had not been informed that they would be tested again on the material that they had covered with the robot. After each class had completed the retention test, the experimenter gave an overview of the study and a presentation of social robots to all children. This meant that all children understood the study and had the opportunity to interact with the robot.

VI. Results

A. Perception of the Robot

To address H1 (that children will perceive differences in the verbal availability of the robot), the results of the post-interaction questionnaire were analysed. The questionnaire is broken down into the several parts which measure different constructs, as described in Section V-B. The manipulations were conducted on the verbal immediacy element of the questionnaire, where a higher verbal immediacy score would indicate a higher perception of verbal availability. An unpaired t-test reveals a significant difference between the average verbal immediacy measure for the LOW condition (M=31.2, 95% CI [28.1,34.3]) and the HIGH condition (M=44.9, 95% CI [41.6,48.2]); $t(38)$=6.322, $p<.001$. This confirms H1; children could indeed perceive the difference between the conditions (despite not having seen the other condition for comparison).

Nonverbal immediacy scores were also compared; the difference between the nonverbal immediacy score in the LOW condition (M=18.5, 95% CI [17.0,19.9]) was not found to be significantly different to that of the HIGH condition (M=19.6, 95% CI [17.8,21.3]); $t(38)$=1.020, p=.314 (Fig. 5). This provides some validation for the control of nonverbal behaviour between the conditions.

B. Learning Gains

Learning gains are measured through scores on the tests conducted before the interaction (pre-test), immediately after the interaction (post-test), and 3-7 days after the interaction (retention test). Questions on the tests are equally weighted, so scores are out of a maximum of 12. Before analysis of the two robot conditions can be conducted there are some potential confounds which must be eliminated as factors: the differences in time between the interaction and retention test, and the impact of exposure to the test (as the same test is used).

It could be expected that children who interacted with the robot at a time closer to the retention test would outperform those who interacted with the robot earlier in the visit. To explore whether this was a factor, the day on which the interaction took place was correlated with the difference between the post-test and the retention test. The correlation is weak and non-significant; $r(36)$=-.079, p=.637, indicating that the time from interaction to retention test can be eliminated as a factor. We would suggest that the absolute number of days does not make a difference to the retention, but the number of days out of school during this period is more important, which was constant for all children (a weekend of 2 days).

The control condition is used to verify whether exposure to the test makes a difference to the findings. It would not be expected that there would be a difference as the children are given no feedback on the tests at any stage, but the control condition allows verification. For children in the control condition, the pre-test score (M=3.96, 95% CI [3.26,4.66]) and retention test score (M=3.89, 95% CI [3.28,4.49]) can be considered equivalent. Two one-sided t-tests (TOST) [24] with a 1 point threshold confirm the test scores are equivalent at the $p<.05$ level: $t(52)=-2.061$, p=.022/$t(52)$=2.391, p=.010. This indicates that exposure to the test is not a confounding factor.

A repeated measures ANOVA was used to explore H2 (that children will retain their learning) and H3 (that the robot condition will affect learning); Fig. 6 and Table I show

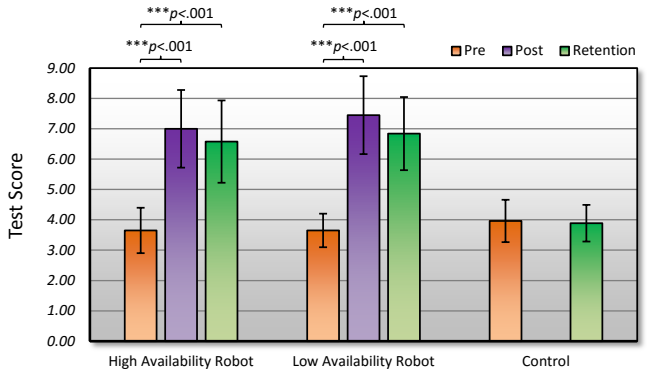


Fig. 6. Pre-test, post-test and retention test scores by condition (chance score=3; maximum score=12). Children learn a significant amount from the robot between pre- and post-tests; this gain is sustained to the retention test. *Error bars* show 95% CI.

the results for test scores by condition. Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated, $\chi^2(2)=1.873$, $p=.392$. No significant interaction was found between test and condition; Wilk's Lambda=.998, $F(2,35)=0.04$, $p=.963$. A main effect was found for test, Wilk's Lambda=.391, $F(2,35)=27.21$, $p<.001$, but not for condition; $F(1,36)=0.08$, $p=.774$. Bonferroni pairwise comparisons find that there is a significant difference between pre-test and post-test, and pre-test and retention test scores (all $p<.001$), but no difference between post-test and retention test ($p=1.00$).

These results support H2, as children learn between the pre- and post-tests, and retain their learning in the retention test. Further support for H2 can be gained through Weber & Popova paired-samples equivalency tests [25] which show that the post and retention test scores are equivalent in both the HIGH ($t(18)=0.67$, $p=.022$) and LOW ($t(18)=0.73$, $p=.025$) conditions, with Cohen's $d=.50$. Whilst this is an 'intermediate' effect size for demonstrating equivalency, it should be noted that the sample size is relatively small on a per-condition basis, leading to a higher variation in scores, which raises the level at which equivalency can be shown. Combined, these findings provide evidence in support of H2 as the children learn a significant amount from the pre-test to the post-test, and the post-test and retention test scores can be considered largely equivalent, demonstrating their retention of the learning.

The ANOVA results do not support H3 (that higher availability will lead to greater learning) as no significant effect was found for robot condition. Nor can a significant difference be seen between the improvement in the LOW condition ($M=3.80$, 95% CI [2.55,5.05]) and the HIGH condition ($M=3.35$, 95% CI [1.78,4.92]); $t(38)=0.470$, $p=.641$. The drop in score from post-test to retention test can also be considered equivalent between conditions; using a Weber & Popova independent-samples equivalance test, $t(36)=0.07$, $p=.004$ with Cohen's $d=.50$. Therefore, Hypothesis H3 must be rejected as there are no significant differences observed between conditions in terms of learning.

Based on the rules taught to the children, one could suggest that learning a very simple rule of: "if the word ends in an 'e', then use *la*, otherwise use *le*" may be adopted as a 'shortcut' and could account for the learning differences. This would then have nothing to do with learning aspects of language, but be a basic memory phenomenon. This had been anticipated in the study design, so later questions in the learning material made sure to challenge this approach by including several words ending in 'e' as possible answers, but with those words relating to humans of male gender (therefore requiring 'le', rather than 'la' and violating the shortcut rule). Additionally, a question in the tests used adopted this approach, with several words ending in an 'e', but not all being feminine. This was done to verify whether the shortcut rule had been adopted, or whether the children had really learnt the material as it had been taught, with the ability to discriminate between different types of words. If the children had only learnt the shortcut rule then they would answer this verification question incorrectly, however, it was answered correctly above the average level for the rest of the questions in the test (63% for the verification question, versus 60% for the other questions). This provides some evidence that the children learnt intricacies of the language that was presented to them; further evidence in support of this will be provided in Section VII.

TABLE I
TEST SCORE RESULTS BY CONDITION.

Condition	Pre-Test *M* [95% CI]	Post-Test *M* [95% CI]	Retention Test *M* [95% CI]
CTRL	3.96 [3.26, 4.66]	N/A	3.89 [3.28, 4.49]
LOW	3.65 [3.10, 4.20]	7.45 [6.17, 8.73]	6.84 [5.64, 8.05]
HIGH	3.65 [2.90, 4.40]	7.00 [5.72, 8.28]	6.58 [5.22, 7.94]

VII. Discussion

The results show that the children perceived the verbal availability of the robot conditions as intended, which confirms that the behaviour was designed appropriately to address the research hypotheses. The nonverbal behaviour was kept constant between the two conditions, and this was reflected in the children's questionnaire responses. The children in both robot conditions exhibited significant learning gains between the pre-test and post-test, as well as between the pre-test and retention test, with equivalent scores in the retention test and the post-test. This is a positive result, as it would have been plausible that the children would quickly forget what the robot had taught them once the interaction was over, especially as the children were not aware that they would be re-tested, and so had little motivation to attempt to actively try and retain the information.

The tests which the children had to complete were designed to be challenging. Each answer had four options with no obviously incorrect answers, so the likelihood of a guess being correct would be chance (25%). It was found that children scored slightly above this on the pre-tests as they had done a small amount of French before, so scored closer to 4 than the 3 that would be expected with random guessing. This significantly improved to over 7 out of 12 in the post-tests.

Given the difficulty of the tests and the relatively short time the child interacts with the robot learning and practising the material, this is an impressive increase. Indeed, only 6 of the 40 children who interacted with the robot did not improve from pre-test to post-test. Learning of 'le' or 'la' as the article choice could have contributed to part of the increase in scores, however if children had learnt the choice to be le/la then the chance score would go up by 1.5 points from pre-test (chance = 3) to post-test (chance = 4.5). The children actually improve by an average of 3.6 (95% CI [2.6,4.5]), suggesting learning beyond any improvement due to the higher chance score.

Despite the children being able to perceive the difference in verbal aspects of availability between the two robot conditions (measured through verbal immediacy), no significant difference was observed in learning in either the post- or retention-test. This finding is surprising given the positive correlation between verbal immediacy and learning in human studies [9], [11]. Previous work has found that nonverbal aspects of availability can lead to additional learning above that gained through just exposure [20]. The work here explored whether verbal aspects of availability would have a similar positive effect on learning, but they did not.

Aspects of the behaviour manipulated here, such as personalisation [12] and off-activity talk [13], have been studied before in HRI with promising results. However, these studies had too few subjects to make conclusions about learning [12], or did not assess learning [13]. In contrast to [13], we do find here that the children perceive differences between the conditions, but in our study the questionnaire is targeted towards specifically measuring the perception of the behaviours which were manipulated, rather than assessing an overall feeling towards the robot. It is possible that despite children perceiving differences in the availability of the robot, this did not translate into any difference in feeling towards the robot. If the relationship the child feels towards the robot is no different between conditions then this may go some way to explaining the lack of difference in learning.

The interpretation of the robot character could have been influenced by the TTS voice used by the robot, which would switch when the language changed. These voices were clearly different and this could have impacted how the children perceived the robot. However, the children have no prior experience with the robot, so they may have accepted this as part of the robot's behaviour. As the voices are clearly different, they may also have interpreted this not to be part of the robot's character, but to be the robot playing back other media (akin to a teacher playing recorded French). It is not possible to determine how the children perceived this switch in voice from the data collected, but perceptions of voice switching of multi-lingual robots could be worth explicitly exploring in future work.

Another factor which may have influenced the learning results is novelty. Novelty is often an issue for HRI studies [26], [27], and it possibly played a role here as the children interact just once with the robot for a brief period of time. Verbal immediacy has been found to consist of four factors, including 'individual friendliness' [10]. Even if the children were to bond more strongly with the high availability robot because of increased friendliness, the short interaction time might not be enough for differences in the relationship to manifest into learning outcomes. Furthermore, it could be that the behaviour of the more available robot cancels out its own benefits by being so novel as to distract from the learning material. For example, when the robot is conducting off-activity talk during the interaction, this is time when the children are not focussing on the learning task and are possibly forgetting information they have learnt. This doesn't mean that off-activity talk should be avoided for fear of distraction, but that it might only be appropriate in longer, or repeated interactions where novelty is less of an issue. We would hypothesise that given a longer interaction timescale, the learning benefits predicted by the literature of greater availability [9], [11] would be observed as the novelty wears off [2], [26].

In the HHI literature, a lower correlation between verbal immediacy and learning has been found when compared to nonverbal immediacy and learning [11]. Nonverbal immediacy has previously been found to make a difference to learning in HRI [20], [21]. This could suggest that verbal behaviour may not be as important for learning (at least in short-term interactions) as overt nonverbal behaviour. It has also been found in humans that the impact of immediacy behaviours is enhanced in line with increases in class size [9]. It could be that the effect of verbal immediacy is simply too far reduced when placed in a one-to-one tutoring context as in this study, rather than the larger classroom setting. The availability of the robot would be experienced to some extent in both conditions simply through the nature of the one-to-one interaction.

One interesting finding from the data collected which was not hypothesised was the ability of the children to acquire vocabulary despite the learning material not explicitly requiring them to do so. Three questions of the test were vocabulary based: two requiring translation from English to French, and one French to English. Two of these questions referred to words which the children would have seen on screen and heard the robot say (as they were answers to questions in the learning material). The remaining question was about a word which they would have seen on screen, but the robot did not say (as it was not a correct answer). It is suggested that the two words which were answers in the learning material were more likely to be recalled as the children would have looked at the word for longer and the robot would have said the word. However, a significant increase was found for all 3 of the questions independently, and a repeated measures ANOVA found a significant increase for the average score (out of 3) of children who correctly translated the words from pre-test to post-test, and from pre-test to retention test. Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated, $\chi^2(2)=0.661$, $p=.719$. No significant interaction was found between test and condition; Wilk's Lambda=.968, $F(2,35)=0.58$, $p=.565$. A main effect was found for test, Wilk's Lambda=.595, $F(2,35)=11.94$, $p<.001$, but not for condition; $F(1,36)=0.14$, $p=.710$. Post-hoc Bonferroni pairwise

comparisons find that there is a significant difference between pre-test (M=0.8, 95% CI [0.6,1.0]) and post-test (M=1.6, 95% CI [1.3,1.9]), and pre-test and retention test (M=1.4, 95% CI [1.1,1.7]) scores ($p<.001$ and $p=.001$, respectively), but no difference between post-test and retention test scores ($p=.883$).

It is of course possible that the children remembered the words from the pre-test and made an effort to learn these words when they were presented on screen, but this seems unlikely given the time (up to 4 days) between many of the pre-tests and the interactions, and the sheer number of words they were exposed to in the learning content (over 40). For a child to concentrate on learning 3 words from the pre-test, days after having seen it, when being taught a different aspect of language would seem to be highly improbable. As such, this is a promising finding with robots that confirms data from human-human literature whereby children of this age will acquire language through exposure in social interactions [14].

VIII. Conclusion

Children perceived the relative social availability of the two robot conditions as intended in the design. This confirms that the manipulations made were appropriate to address the question of whether an increase in verbal aspects of availability would lead to an increase in learning. As expected, the children did learn elements of a second language from the robot. This was measured immediately after the interaction and also some days later. The retention test scores were slightly lower than the pre-test scores, but can be considered statistically equivalent. However, surprisingly there was a lack of any significant difference between conditions in the immediate post-test score, or the longer-term retention test score. Literature from human-human interaction studies [9], [11] and human-robot interaction studies [12], [13] would predict an increase in robot verbal availability to lead to an increase in learning, but this was not found. These findings suggest that in this short-term dyadic interaction context, additional effort in developing social aspects of a robot's verbal behaviour may not return the desired positive impact on learning gains.

IX. Acknowledgements

This work is funded by the EU FP7 DREAM project (grant 611391), H2020 L2TOR project (grant 688014), and SoCEM, Plymouth University, U.K. Thanks goes to Dr. Caroline Floccia who provided valuable feedback on the study design.

References

[1] M. Alemi *et al.*, "Employing Humanoid Robots for Teaching English Language in Iranian Junior High-Schools," *Int. Journal of Humanoid Robotics*, vol. 11, no. 3, 2014.

[2] T. Kanda *et al.*, "Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial," *Human-Computer Interaction*, vol. 19, no. 1, pp. 61–84, 2004.

[3] E. Short *et al.*, "How to Train Your DragonBot: Socially Assistive Robots for Teaching Children About Nutrition Through Play," in *Proc. of the 23rd IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2014, pp. 924–929.

[4] G. Gordon *et al.*, "Can Children Catch Curiosity from a Social Robot?" in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI*. ACM, 2015, pp. 91–98.

[5] J. Kennedy *et al.*, "The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning," in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI*. ACM, 2015, pp. 67–74.

[6] ——, "Can Less be More? The Impact of Robot Social Behaviour on Human Learning," in *Proc. of the 4th Int. Symp. on New Frontiers in HRI at AISB 2015*, 2015.

[7] P. K. Kuhl, "Cracking the speech code: How infants learn language," *Acoustical Science and Technology*, vol. 28, no. 2, pp. 71–83, 2007.

[8] J. Herberg *et al.*, "Robot watchfulness hinders learning performance," in *Proc. of the 24th IEEE Int. Symp. on Robot and Human Interactive Communication*, 2015.

[9] J. Gorham, "The relationship between verbal teacher immediacy behaviors and student learning," *Communication education*, vol. 37, no. 1, pp. 40–53, 1988.

[10] J. H. Wilson and L. Locker Jr, "Immediacy scale represents four factors: Nonverbal and verbal components predict student outcomes," *The Journal of Classroom Interaction*, vol. 42, no. 2, pp. 4–10, 2007.

[11] P. L. Witt *et al.*, "A Meta-Analytical Review of the Relationship Between Teacher Immediacy and Student Learning," *Communication Monographs*, vol. 71, no. 2, pp. 184–207, 2004.

[12] O. A. Blanson Henkemans *et al.*, "Using a robot to personalise health education for children with diabetes type 1: A pilot study," *Patient Education and Counseling*, vol. 92, no. 2, pp. 174–181, 2013.

[13] I. Kruijff-Korbayova *et al.*, "Effects of Off-Activity Talk in Human-Robot Interaction with Diabetic Children," in *The 23rd IEEE Int. Symp. on Robot and Human Interactive Communication*, 2014, pp. 649–654.

[14] P. K. Kuhl, "Brain mechanisms in early language acquisition," *Neuron*, vol. 67, no. 5, pp. 713–727, 2010.

[15] J. K. Westlund and C. Breazeal, "The interplay of robot language level with children's language learning during storytelling," in *Proc. of the 10th ACM/IEEE Int. Conf. on HRI Extended Abstracts*. ACM, 2015, pp. 65–66.

[16] M. Saerbeck *et al.*, "Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, 2010, pp. 1613–1622.

[17] V. P. Richmond *et al.*, "Development of the Nonverbal Immediacy Scale (NIS): Measures of Self- and Other-Perceived Nonverbal Immediacy," *Communication Quarterly*, vol. 51, no. 4, pp. 504–517, 2003.

[18] A. Mehrabian, "Some Referents and Measures of Nonverbal Behavior," *Behavior Research Methods & Instrumentation*, vol. 1, no. 6, pp. 203–207, 1968.

[19] T. Belpaeme *et al.*, "Multimodal Child-Robot Interaction: Building Social Bonds," *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 33–53, 2012.

[20] J. Kennedy *et al.*, "Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions," in *Proc. of the Int. Conf. on Social Robotics*, 2015, pp. 327–336.

[21] D. Szafir and B. Mutlu, "Pay Attention!: Designing Adaptive Agents that Monitor and Improve User Engagement," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, 2012, pp. 11–20.

[22] F. Tanaka and S. Matsuzoe, "Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning," *Journal of Human-Robot Interaction*, vol. 1, no. 1, pp. 78–95, 2012.

[23] K. Board and T. Tinsley, *Language Trends 2014/15: The state of language learning in primary and secondary schools in England*. CfBT Education Trust, 2015.

[24] D. J. Schuirmann, "A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability," *Journal of pharmacokinetics and biopharmaceutics*, vol. 15, no. 6, pp. 657–680, 1987.

[25] R. Weber and L. Popova, "Testing equivalence in communication research: Theory and application," *Communication Methods and Measures*, vol. 6, no. 3, pp. 190–213, 2012.

[26] I. Leite *et al.*, "Social robots for long-term interaction: A survey," *Int. Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013.

[27] J. Sung *et al.*, "Robots in the wild: understanding long-term use," in *Proc. of the 4th ACM/IEEE Int. Conf. on HRI*. IEEE, 2009, pp. 45–52.

4 2015 Publications

- Tony Belpaeme, James Kennedy, Paul Baxter, Paul Vogt, Emiel J. Krahmer, Stefan Kopp, Kirsten Bergmann, Paul Leseman, Aylin C. Küntay, Tilbe Göksun, Amit K. Pandey, Rodolphe Gelin, Petra Koudelkova, and Tommy Deblieck (2015) L2TOR – Second Language Tutoring using Social Robots. In *Proceedings of 1st Int. Workshop on Educational Robots.* Springer.

L2TOR - Second Language Tutoring using Social Robots

Tony Belpaeme[1], James Kennedy[1], Paul Baxter[1], Paul Vogt[2], Emiel E.J. Krahmer[2], Stefan Kopp[3], Kirsten Bergmann[3], Paul Leseman[4], Aylin C. Küntay[5], Tilbe Göksun[5], Amit K. Pandey[6], Rodolphe Gelin[6], Petra Koudelkova[6], Tommy Deblieck[7]

[1] Plymouth University, UK, [2] Tilburg University, The Netherlands, [3] University of Bielefeld, Germany, [4] University of Utrecht, The Netherlands, [5] Koç University, Turkey, [6] Aldebaran Robotics, France and [7] QBMT, Belgium

Abstract. This paper introduces a research effort to develop and evaluate social robots for second language tutoring in early childhood. The L2TOR project will capitalise on recent observations in which social robots have been shown to have marked benefits over screen-based technologies in education, both in terms of learning outcomes and motivation. As language acquisition benefits from early, personalised and interactive tutoring, current language tutoring delivery is often ill-equipped to deal with this: classroom resources are at present inadequate to offer one-to-one tutoring with (near) native speakers in educational and home contexts. L2TOR will address this by furthering the science and technology of language tutoring robots. This document describes the main research strands and expected outcomes of the project.

1 Background

Second language learning has become an important element of formal education for many children in Europe and beyond. For some children, the language used at school is a second language (noted as L2), as they speak a different language or dialect at home. This not only holds for immigrant children, but also for children speaking an official minority language of their country of residence. Preschool years are important to develop adequate knowledge of the academic language, as later educational success builds on it (Leseman & van den Boom, 1999; Hoff, 2013). Thus, it is essential that children with a different home language than the dominant one receive "sensitive" bilingual input and interaction once they enter day care and preschool settings. The robot tutor we propose here serves that crucial aim.

The current challenges of standard L2 teaching in classrooms are that the interaction between tutors and students often is one-to-many. In addition, language teaching does not reflect how language is naturally acquired and the tutor is often either not fluent in the second language or not proficient in the child's mother tongue. While there is large variation in L2 proficiency in young children, with factors such as gender, socio-economic background and home education having a significant impact, there is ample evidence for the current language education provision and the young learners' subsequent L2 performance being on occasion suboptimal (Brühwiler and Blatchford, 2011; De Feyter and Winsler, 2009; Kim et al., 2014). While a number of educational approaches

adfa, p. 1, 2011.

remedy this through, for example, immersion approaches, second language teaching remains challenging, especially for immigrant children (Collins et al., 2012).

It has long been established that one-to-one tutoring can result in significantly higher cognitive learning gains than group education. Bloom (1984) found that one-to-one tutoring resulted in 2 standard deviations improvement against a control group, concluding that "the average tutored student was above 98% of the students in the control class" (p. 4). Whilst research since has shown that the effects are not as large as first observed, there is nonetheless a distinct advantage to the one-to-one tutoring approach (VanLehn, 2011). However, traditional school classroom arrangements mean that one teacher is responsible for many children. In such situations it is not possible for teachers to offer as much one-to-one tutoring as would be desired.

More recently it has emerged that social robots can be used to teach children and adults. However, what is remarkable is that social robots seem to have a distinct advantage over alternative digital one-to-one tutoring technologies, such as screens and tablets. When tutoring is delivered by a social robot this leads to greater learning gains compared to the same content delivered on-screen (Han et al., 2005, Hyun et al., 2008, Kose-Bagci et al., 2009, Leyzberg et al., 2012), with performance increases of up to 50% compared to interactive screen technology (Kennedy et al. 2015). The reasons for this are still unclear: it might be that the social and physical presence of the robot engages the learner more than just on-screen delivery and feedback, or it might be that the learning experience is a more multimodal experience thus resulting in a richer and embodied pedagogical exchange (Mayer & DaPra, 2012), or of course a combination of these two.

Of importance here is that robots (and digital media such as tablets and computers) allow for a fast-paced interaction, and digital devices can tailor the interaction to match the level and interests of the young learner. This allows for the system to stay within Vygotsky's Zone of Proximal Development of the child and adopts an interactionist perspective to learning (Chapman, 2000); both approaches are central to this project.

L2TOR (pronounced 'el tutor'), runs for 3 years starting early 2016, and aims to design a child-friendly tutor robot that can be used to support teaching preschool children a second language by interacting with children in their social and referential world. In particular, the project will focus on teaching English as L2 to native speakers of Dutch, German and Turkish, and teaching Dutch and German as L2 to immigrant children speaking Turkish as a native language. The L2TOR robot will be designed to interact naturally with children aged four years old in both the second language and the child's native language. The robot's social behaviour will be based on how human tutors interact with children, and will not only use verbal communication, but also non-verbal communication, such as gestures and other forms of body language. The robot will be able to adaptively respond to children's actions and engage with them in tutoring interactions. The child will be provided with increasingly complex stimuli and utterances in the second language, as well as appropriate feedback that support the child's language development.

2 General approach

The central goal of the L2TOR project is to develop an embodied digital learning environment in which a child-friendly, social humanoid robot serves as a tutor to assist children acquiring a second language. This robot will be able to interact with the child naturally at a level that challenges the child to learn new words and grammar, while at the same time feels like a friend. The robot will keep track of individual children's development and will adapt its own interaction to facilitate the child to advance to the next level. As such, the robot will construct a scaffold that allows the child to acquire new skills in interaction. Since the robot will teach the child a second language, proficiency in the child's native language is desirable, so it can provide explanations and instructions that the child can readily understand.

The L2TOR embodied digital learning environment will not only consist of the robot, but it is a complete learning environment that also consists of a table-top environment that represents the contextual content of the system. Depending on the educational domain, this table-top environment will either be a table with moveable objects or an interactive tablet computer (Fig 1). Together with the child, the robot and table-top will constitute the contextual setting in which the tutoring will take place.

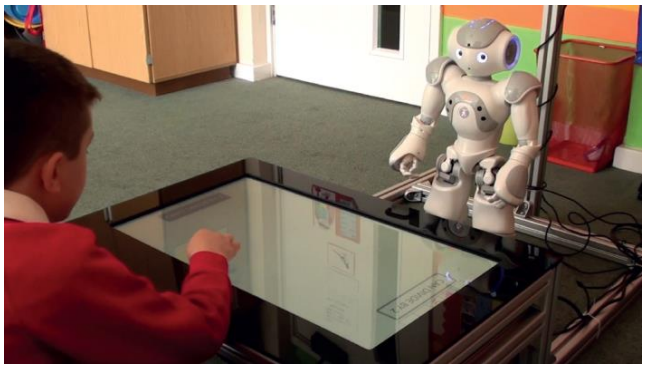


Fig. 1. A robot teaching division skills and prime numbers to a primary school pupil (Kennedy et al., 2015). L2TOR will use a similar setup, using a tablet computer instead of a larger display.

To develop an effective tutoring robot, the robot should interact with a child in similar ways a caregiver or teacher would do when teaching the child language. Such interactions not only include verbal content, but also nonverbal content and adequate socio-cognitive skills, because these form the pragmatic backbone of language acquisition from infancy on (Matthews, 2014). This multimodal interaction allows the interactants to construct and maintain common ground, which is essential for language learning, because this provides the child with a suitable context to learn from. Since there are few observational data on multimodal interaction for L2 language tutoring, we will collect our own data and analyse these such that they can be incorporated as a template for the L2TOR robot.

The primary requirement for building common ground is to design child-robot interactions that allow for mutual understanding of the communicative acts and the environment in which the interactants are situated. For the L2TOR robot, this means that the robot should be able to

- perceive and recognize the objects and events that occur in the environment,
- perceive and recognize the verbal and nonverbal signals produced by the child,
- use Theory of Mind to take the child's perspective,
- be able to monitor linguistic/behavioural errors produced by the child,
- respond to the child in a contingent manner, both temporally and semantically and,
- produce appropriate utterances in different modalities (particularly, speech and gesture) and in different languages (native and target language).

The design of these capacities is the major challenge that the L2TOR project needs to tackle, but current technology is sufficiently advanced to provide pragmatic solutions for most issues. For example, the perception and recognition of social signals is unsolved for open domains, but early work shows that for closed-domain interactions, we have sufficient interpretability to allow for full autonomy (Kennedy, Baxter & Belpaeme, 2015). As far as possible, the implementation will rely on integrating existing technologies, especially for the hardware solutions, the input recognition and the motor control of the system. A key point here will be speech recognition, with current speech recognition system not performing with sufficient reliability for child speech; to mitigate this, interaction which be directed through a touch screen interface on which the young learner taps icons. The tutor robot will be realised by Aldebaran Robotics' Nao humanoid (Fig. 1), which comes with a large range of suitable software for input and output processing. The challenges occur in the design and implementation of multimodal interactions that have the capacity construct common ground with the child to facilitate L2 acquisition.

3 Three lesson series

While interaction design for robots has been explored extensively, research into how interactions should be designed to support tutoring and teaching is recent and as of yet inconclusive. As a first goal, the pedagogy of robot assisted language tutoring will have to be defined. For this purpose, the L2TOR project will design, implement and evaluate three series of lessons (each running 10-15 weeks, 3-4 sessions per week) for the three educational domains:

1. Number domain: Learning language about basic number and pre-mathematical concepts.
2. Space domain: Learning language about basic spatial relations.
3. Storytelling domain: Vocabulary and concept learning during storytelling.

These domains were chosen to restrict the range of interactions such that the objectives are feasible and measurable within the duration of the project, while at the same time being relevant and suitable for educational purposes in a pre-school setting. Each

lesson will be implemented and evaluated for five language pairs L1 and L2: native speakers of German, Dutch and Turkish will be taught English, while Turkish (immigrant) children will be taught Dutch or German, depending on their country of residence. These language combinations are not only chosen for practical considerations (they cover the languages of the academic partner states involved), but also for strategic reasons. First, English is the most commonly taught second language across Europe. Second, many children of Turkish immigrants live in the Netherlands and Germany, and will learn Dutch or German in preschool and beyond. Thus, the latter will represent a common situation of ethnic minority children learning L2 at school.

For each domain, learning targets will be developed. In the number domain, the learning targets will increase in complexity from mere counting objects and naming of shapes, to comparing numerosities, and to performing transformations on objects and sets (addition, subtraction, identifying geometrical shapes). In the space domain, learning targets range from exploring spatial relations between objects from an egocentric perspective (preposition and movement verbs), to spatial relations from an allocentric perspective (navigation through space and perspective taking), and performing a construction task (building a model with blocks) following instructions involving spatial relations, spatial coordinates and movement through space. The learning targets for the storytelling domain include vocabulary about rare objects and events (e.g., "wooden bird", "magical flying bird"), and basic narrative structures.

For each lesson series, the L2TOR robot will communicate with the child following a specified scenario to obtain the learning targets. These scenarios describe the general sequence of targets that L2TOR aims to achieve by interacting with the child. The scenarios need to be adaptive, because the interactions between child and robot are adaptive and to some extent unpredictable. The contexts for the number and space domains are provided by a blocks/toy world that the children and -to a limited extent- the robot can manipulate. For the number domain, scenarios will be designed in which the objects can be grouped in countable sizes. For the space domain, blocks can be positioned in different ways (e.g., putting blue block on the red block) to test children's use of spatial language for spatial relations between objects. In the storytelling domain, the L2TOR will show the child on the tablet a story about a (currently not available) "magical transformation machine", where a character (e.g. a wooden bird) chooses an object among several objects, puts is through a device and transforms into another object (e.g. a flying animate bird). The children will first be asked to form narratives about what they have watched. Later, the child will be given the opportunity to join in a different version of this story with the characters and actions of her own choice.

4 References

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*: 4-16.

Brühwiler, C., & Blatchford, P. (2011). Effects of class size and adaptive teaching competency on classroom processes and academic outcome. *Learning and Instruction*, 21(1):95-108.

Chapman, R. S. (2000). Children's language learning: An interactionist perspective. *Journal of Child Psychology and Psychiatry*, 41(1):33-54.

Collins, M.F. (2010). ELL preschoolers' English vocabulary acquisition from story-book reading. *Early Childhood Research Quarterly*, 25, 84–97.

De Feyter, J. J., & Winsler, A. (2009). The early developmental competencies and school readiness of low-income, immigrant children: Influences of generation, race/ethnicity, and national origins. *Early Childhood Research Quarterly*, 24(4):411-431.

Han, J., Jo, M., Park, S., and Kim, S. (2005). The educational use of home robots for children. In *Proceedings of the 14th IEEE International Symposium on Robots and Human Interactive Communications, RO-MAN 2005*, pages 378-383. IEEE.

Hoff, E. (2013). Interpreting the Early Language Trajectories of Children from Low SES and Language Minority Homes: Implications for Closing Achievement Gaps. *Developmental Psychology*, 49(1):4–14.

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 Robots and Human Interactive Communications, RO-MAN 2008*, pages 187-192. IEEE.

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 *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, pages 67-74. ACM.

Kim, Y. K., Curby, T. W., & Winsler, A. (2014). Child, family, and school characteristics related to English proficiency development among low-income, dual language learners. *Developmental psychology*, 50(12):2600.

Kose-Bagci, H., Ferrari, E., Dautenhahn, K., Syrdal, D. S., and Nehaniv, C. L. (2009). Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*, 23(14):1951-1996.

Leseman, P. P. M., & van den Boom, D. C. (1999). Effects of quantity and quality of home proximal processes on Dutch, Surinamese–Dutch and Turkish–Dutch preschoolers' cognitive development. *Infant and Child Development*, 8(1):19-38.

Leyzberg, D., Spaulding, S., Toneva, M., and Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society, CogSci 2012*, pages 1882-1887.

Matthews, D. (Ed.). (2014). Pragmatic development in first language acquisition (Vol. 10). John Benjamins Publishing Company.

Mayer, R. E., & DaPra, C. S. (2012). An Embodiment Effect in Computer-Based Learning With Animated Pedagogical Agents. *Journal of Experimental Psychology*: *Applied*, 18(3), 239–252.

VanLehn, K. (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197-221.