Varied Human-Like Gestures for Social Robots: Investigating the Effects on Children's Engagement and Language Learning

Jan de Wit
Tilburg University
Tilburg, the Netherlands
j.m.s.dewit@uvt.nl

Arold Brandse
Tilburg University
Tilburg, the Netherlands
m.a.j.brandse@uvt.nl

Emiel Krahmer
Tilburg University
Tilburg, the Netherlands
e.j.krahmer@uvt.nl

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

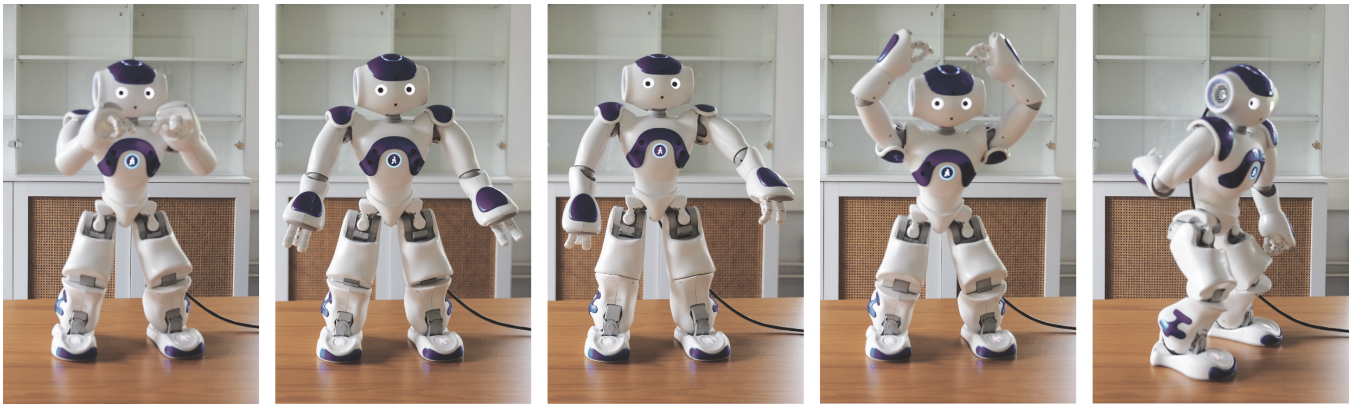

Figure 1: We investigated whether a robot can use iconic gestures to support its teaching activities, and if it helps to add variation to these gestures. These are the gesture variations for *turtle*. Videos available at https://tiu.nu/hri20-gestures

ABSTRACT

To investigate whether a humanoid robot's use of gestures improves children's learning of second language vocabulary, and if variation in gestures strengthens this effect, we conducted a field study where a total of 94 children (aged 4–6 years old) played a language learning game with a NAO robot. The robot either used no gestures at all, repeated the same gesture every time a target word was presented, or produced a different gesture for each occurrence of a target word. We found that, contrary to what the majority of existing research suggests, the robot's use of gestures did not result in increased learning outcomes, compared to a robot that did not use gestures. However, engagement between child and robot was higher in both the repeated and varied gesture conditions, compared to the condition without gestures. An exploratory analysis showed that age played a role: the older children in the study learned more than the younger children when the robot used gestures. It is therefore important to carefully consider the design and application of robot gestures to support the learning process. The contribution of this work is twofold: it is a conceptual reproduction of a previous study, and we have taken first steps towards exploring the role of variation in gestures. The study was preregistered, and all materials are made publicly available.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**;
• **Applied computing** → **Interactive learning environments**.

KEYWORDS

nonverbal communication, social robotics, robot tutoring, second language learning, human-robot interaction

ACM Reference Format:
Jan de Wit, Arold Brandse, Emiel Krahmer, and Paul Vogt. 2020. Varied Human-Like Gestures for Social Robots: Investigating the Effects on Children's Engagement and Language Learning. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), March 23–26, 2020, Cambridge, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3319502.3374815

HRI '20, March 23–26, 2020, Cambridge, United Kingdom

ACM ISBN 978-1-4503-6746-2/20/03…$15.00
https://doi.org/10.1145/3319502.3374815

1 INTRODUCTION

Manual gestures [18] are an essential part of our everyday communication with other people: we produce them naturally to support our thinking process, and use them to avoid miscommunication [14]. Specifically iconic gestures — a subset of gestures where the movements are meaningfully linked to the concept that is referred to [27] — are known to be a valuable support mechanism in education, resulting in improved learning outcomes and higher levels of engagement from the student with the educational process [17, 41, 48]. The present work focuses on the domain of second language (L2) learning, where gestures have been shown to contribute to increased vocabulary acquisition [12, 23, 40]. They enable "grounding" of new knowledge in existing sensorimotor experiences [4]. For example, when teaching students about the word *ball* in a second language, by accompanying this unknown word with an iconic gesture depicting the underlying concept of a ball (e.g., by molding its shape or by bouncing an imaginary ball) we provide additional scaffolding to create a link between the new form (the L2 word) and the learner's existing knowledge of its corresponding meaning.

With an increasing research interest into using robots in contexts where they are expected to interact socially with humans, such as education [5] and specifically second language learning [16, 42], a number of groups have started exploring whether gestures result in similar positive effects when they are being performed by a robot instead of a human. A survey comparing robots to virtual agents indicates the robot's ability to move and perform gestures in the physical world to be one of its key advantages over screen-based alternatives [22]. Observed effects of a robot's use of gestures include increased memorization of story details by the listener [7, 15], better human-robot collaborative task performance [6], and higher levels of engagement with the robot [7, 11, 35]. Furthermore, a robot that gestures is generally perceived more positively [1, 2, 33], especially when its motions are exaggerated and cartoon-like [11].

However, applications of robots that gesture in an educational context, and specifically in (second) language learning, remain underresearched. One example can be found in first language learning with adults, where participants that interacted with a robot that used iconic gestures had better learning outcomes than those that did not receive support from the robot's gestures [43]. We recently conducted two studies in second language learning, both with young children as participants, with mixed results. A first exploration showed increased vocabulary retention over time, as well as higher levels of engagement with the robot for children that received additional support from the robot in the form of iconic gestures compared to children that were not presented with gestures [8]. Our second study, although similar in design, did not see such an effect on learning outcomes [47]. The two studies differed in their duration (number of sessions) and the vocabulary that was taught. The first study consisted of only one session, and taught six animal names while the second study was longitudinal (seven sessions) and contained a larger set of more abstract target words (e.g., *more* and *above*), potentially leading to a lower degree of iconicity in the gestures. In view of these conflicting findings, we set out to test the effects of gestures in a single lesson of second language learning with a robot, but using a more diverse set of target words. This part of the current study is a conceptual replication [49] of our previous work [8] as it includes two of the same experimental conditions from the original study — one where the robot uses iconic gestures, and one where the robot does not use any gestures — although with different target words and several improvements to the measurements. It is important to highlight that the design of robot gestures in earlier studies often only relied on the intuitions of the researchers. Here, instead of defining and designing gestures ourselves, we looked at existing sources to see how humans depict the words. We expected to find results that match those found in the original study, leading to the following two hypotheses:

- H1 Children will learn more target words in a second language (H1a) and remember them better (H1b) when a robot produces iconic gestures for the target words than when the robot does not produce such gestures.
- H2 Children are more engaged when interacting with a robot that produces iconic gestures for the target words than with a robot that does not produce such gestures.

When we gesture, we choose which aspect of a concept to describe with our movements, and which strategy — or mode of representation [28] — we use to depict it: do we focus on shape, such as the roundness of a ball, or rather the act of throwing or shooting a ball? Although people generally have default strategies, there is still a degree of individual variation in how we produce gestures [26, 30, 44, 45]. This variation can partly be explained by cultural differences [19], as well as age [25, 34, 38]. Children tend to maintain a smaller symbolic distance to the concept they are describing, which means they will often use a larger gesture space (body parts or the full body), while adults tend to take the "outsider looking in" perspective, and use only their hands to represent objects or characters [34]. For example, when depicting a *pencil* children are more likely to raise their hands above their head in a pointy shape, representing the pencil with their entire body, compared to adults who generally use their hands to display the act of writing, or outline the shape of a pencil.

How we represent concepts in gesture might also be related to what Piaget defined as schemata, mental representations describing the objects and concepts we know, and any past experiences or actions related to these objects [31]. For example, our schema of a toothbrush could include some of its typical visual features, as well as the act of brushing our teeth. As we develop and experience more aspects of a particular concept, our schema of this concept becomes more elaborate. A related framework is variation theory, which states that the object of learning (e.g., in our case the concepts to which we want to link L2 words) may be perceived differently between people, where one learner might focus on different aspects than another [24]. This theory suggests to add variation to learning examples, thus highlighting multiple features of the object of learning. Both theories identify a certain amount of pre-existing knowledge in the learner — which varies between individuals, and grows with experience — to which new features can be added [13]. This, combined with the fact that we use different strategies for producing gestures, raises the question whether we also have different preferences and skills when it comes to understanding and integrating gestures.

There appears to be no existing research that looked into possible benefits of using variation in gestures to support learning. However,

there have been studies in the context of second language learning where variations were introduced in the number of different speakers [3], reporting better learning outcomes compared to the use of a single speaker. Another study varied the images that were used to support second language learning [36]. Contrary to what was found with variations in speech, this had an adverse effect on the number of newly acquired vocabulary items compared to repeating the same image. The researchers suggest that this may have been caused by shifting the focus from the form (the L2 word and how it is pronounced — the new knowledge that is being taught) to the existing meaning assigned to it by the learner (represented by the image). In the present study we have kept both speech and supporting imagery constant throughout the interaction, while variation is added to the additional gesture modality.

Based on the aforementioned theories, we hypothesize that variation in the robot's gestures results in a greater chance that the gestures align with existing salient features of the underlying concept that are already part of the learner's schemata. Furthermore, by presenting several different features the learner might create a stronger link between the word and the underlying concept, rather than merely linking words to specific stimuli. Existing research also indicates that children are more engaged when interacting with robots that show less repetitive behavior [39]. We therefore hypothesize:

- H3 Children will learn more target words in a second language (H3a) and remember them better (H3b) when a robot produces a different iconic gesture every time a particular target word is presented than when the robot produces the same iconic gesture every time a target word is presented.
- H4 Children are more engaged when interacting with a robot that produces a different iconic gesture every time a particular target word is presented than with a robot that produces the same iconic gesture every time a target word is presented.

Because the ability to interpret gestures grows with age [29, 37], we also explore whether differences in age within our participant group have affected their learning outcomes or engagement. The present study adds to existing research in the field of human-robot interaction and gesture studies by verifying whether the previously observed positive effects of gestures persist when the concepts that are taught are more diverse. Furthermore, we investigate whether the previously unresearched addition of variation in a robot's repertoire of gestures further increases these effects. We also propose several improvements to the process of measuring learning outcomes and engagement, with the goal of improving the reliability of our findings. Our hypotheses and planned statistical analyses were preregistered[1], and all of the source code and materials needed to replicate this study are made publicly available[2].

2 DESIGN OF THE INTERACTION

We used the one-on-one tutoring interaction from our previous study [8], in which a child and a SoftBank Robotics NAO robot together played a simplified version of the game *I spy with my little eye*, which is described in more detail below. Two minor changes were made to the original source code. First, the target words were changed to include a more diverse set of objects: bridge, horse, pencil, spoon, stairs, and turtle. Second, we implemented the additional experimental condition in which the robot used a different gesture every time a target word was presented. The five available gestures for each concept were randomized for each participant, so that no order effects could occur. We now briefly explain the process of designing and validating the gestures, and the workings of the educational game that was used.

2.1 Gestures

In order to ensure that only gestures that participants were likely to recognize were used, all of the robot's depictions were based on an existing dataset of recordings from humans producing silent gestures in the context of a game of charades with a robot [9]. We based our choice of target words on the availability of varied examples within this dataset, while ensuring that they covered a diverse range of categories (e.g., tools, static objects, animate objects). We also took into account the age of acquisition [21] for the words, so that the children in our study should know them in their first language. Although the dataset includes three-dimensional Kinect recordings, directly mapping those onto the NAO robot resulted in noisy and unclear gestures. We therefore recreated them by defining keyframes using the Choregraphe software that is distributed with the NAO robot [32], while staying true to the recorded gestures as much as possible. This is a common workflow for creating robot motion that was also used in the original study [8], but now based on examples of people performing the gestures rather than the researchers' frame of reference. Out of the 30 gestures that were implemented, 16 were based on recordings from male performers and 14 from females. Nineteen gestures were recorded from primary school-aged children (6–12 years old), another 10 by adults (20–62 years old), and 1 by a teenager (15 years old).

After recreating the gestures, we video recorded the robot as it performed them and evaluated their clarity by means of an online questionnaire. A total of 19 participants (10 male and 9 female, $M_{age} = 38$ years, $SD = 15$ years) was recruited through convenience sampling. They were shown a video of a gesture and were asked to select the matching concept out of all six included in the study, to investigate whether the gestures were unique enough within the set of six target words. Out of the 30 gestures, 8 scored poorly (< 60% accuracy), 9 scored moderately (60–70%), and 13 scored strongly (> 70%). Based on these findings and additional qualitative feedback, 14 of the gestures were revised to more closely match the human-performed examples from the dataset. Figure 1 shows the five variations for the target word *turtle*. For the experimental condition where the robot did not vary its gestures, we implemented the example that scored highest in the questionnaire (the middle image in Figure 1 for *turtle*).

2.2 Language Learning Game

To train the six target words in the L2 (English), the child and the robot engaged in a simplified version of the game *I spy with my little eye*. The set-up of the experiment included the robot, and a tablet on which the child was able to select answers (see Figure 2). During the training the child sat at a table on which the tablet was placed at a slightly tilted angle. The robot was standing opposite

[1] https://aspredicted.org/wj24k.pdf
[2] https://github.com/l2tor/animalexperiment/tree/variation

Figure 2: The set-up of the experiment at one of the schools.

the child and was put in breathing mode, meaning that it moved its head and arms around slightly and shifted its weight between its legs in order to appear more lifelike.

The robot started by greeting the child with his/her name and then explaining the game, after which the child was asked to indicate whether he or she understood the instructions by touching either a green or red smiley face on the tablet. If the child did not understand the concept of the game, a researcher stepped in to provide further explanation. The game then started with two practice rounds, which were always for the target word *horse* — one in the first language, or L1, Dutch and one in the L2, English — followed by 30 rounds of the game. Each round started with the robot calling out a target word: "Ik zie ik zie wat jij niet ziet, en het is een... *horse*" ("I spy with my little eye a... *horse*"). Three images then appeared on the tablet screen: the correct answer, along with two randomly chosen distractor images (Figure 3). Three images were shown to ensure that the difficulty level while children were still learning was lower than during the post-tests (with six images). The robot provided feedback in response to the child's answer, in which the L2 target word was mentioned again but without any gestures. If the child selected the wrong image, a "repair round" took place where the robot spied the same word once more, but now only the correct image and one distractor image — the previously given answer — were shown.

During the 30 rounds, each of the six target words was presented five times in total, but their order was randomized. In the experimental condition with repeated gestures, the same gesture was used for all five times each target word was presented. In the condition with variation in gestures, the target word in every round was accompanied by a different gesture for that word, but for repair rounds

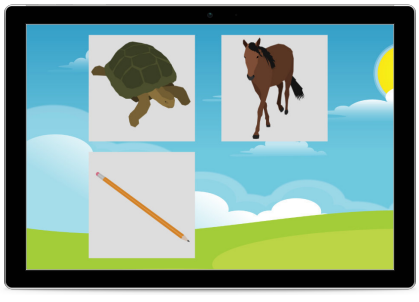

Figure 3: Children provided answers on a tablet screen.

the same gesture from the main round was used. The condition without gestures was identical to the others, but no gestures were used at all. After finishing all 30 rounds, the robot said goodbye to the child. The researcher had a control panel where the child's name was entered, which was used by the robot to personalize the introduction. After pressing a *Start* button, the robot operated fully autonomously, but the interaction could be paused at any time by the researcher if a break was needed. Autonomous behavior was possible by minimizing the complexity of the interactions — the robot did not "listen" to the child, answers to its questions were given through the tablet device.

3 METHODOLOGY

In order to investigate whether the robot's use of iconic gestures resulted in increased learning outcomes and higher levels of learner engagement compared to a robot that does not use such gestures, and to see whether variation in gestures increases learning outcomes and engagement more than repeating the same gesture, we conducted an experiment with the following three experimental conditions: (1) No gestures, where no iconic gestures were included at all; (2) Repeated gestures, where the robot used the same gesture every time a target word came up in the game; (3) Varied gestures, where the robot used five different gestures — a new one for every time a target word came up in the game. Other than these differences in the robot's use of gestures, the experimental conditions were identical, and all children engaged in the same previously described language learning game.

3.1 Participants

A total number of 116 children, recruited from two different primary schools in the Netherlands participated in the study. However, 22 participants had to be excluded due to technical or procedural issues ($N = 12$), bilingualism ($N = 3$), English pre-test scores that were too high (more than four out of six correct, $N = 3$), and missing results due to drop-out ($N = 4$). As a result, the data of 94 children were included in our analyses. The participants were pseudo-randomly assigned to one of the three conditions with a balanced distribution of age and gender (see Table 1 for demographic information). The study was approved by the research ethics committee of Tilburg University. Informed consent was given by the parents of the children prior to their participation.

3.2 Pre-Test and Post-Tests

Children's vocabulary knowledge was measured at different times by means of a test, where images for all six target words were presented on a laptop screen (Figure 4). A voice recording then

Table 1: Demographic Information of Study Participants

Experimental condition	N	Age (Y;M) $\pm SD(M)$	Boys/girls
No gestures	33	5;3 ±9	51% / 49%
Repeated gestures	32	5;2 ±9	56% / 44%
Varied gestures	29	5;4 ±8	41% / 59%
Total	94	5;3 ±9	50% / 50%

asked the child to identify the matching image for a particular target word: "Waar zie je een... [word]?" ("Where do you see a... [word]?"). To reduce bias due to random guessing, in the L2 version each target word was tested three times, yielding a total of 18 test items. To ensure that the test also measured generalizable knowledge, such that the L2 words were not simply linked to the images as they came up in the training session with the robot but rather to the underlying concepts, each of the three times a different image was used: either the same image from training, a photorealistic version, or a line drawing. A target word was scored as correct if the child managed to identify it correctly in at least two out of the three rounds, resulting in a final score of 0–6. In the L1 version of the test each target word was tested only once to save time, and because we assumed that children already knew all of the words in their first language.

3.3 Procedure

3.3.1 Group Introduction. Based on our previous experience working with children and robots, as well as reports from other studies [10, 46], we organized a group introduction to help the children feel at ease with the robot. This was done for entire classrooms at the same time, with the teacher also present. In this session the researchers introduced the robot and demonstrated some of its features. Children were then allowed to shake hands with the robot and put it to bed. The introduction took approximately 15 minutes.

3.3.2 Pre-Test. To measure the pre-existing knowledge of the target words in the L1 and L2, each child was retrieved from the classroom and was asked to complete the test on the laptop, as previously described in Section 3.2. The pre-tests were planned on the same day as the group introduction or shortly thereafter, without the robot present. The tests took approximately 10 minutes and included additional questions related to the children's perception of the robot which are not further analysed here (and were not part of the preregistered analyses).

3.3.3 Training and Immediate Post-Test. The actual training session was scheduled at least one day after the pre-test. The child was retrieved from the classroom and brought to the experiment room. This session consisted of three parts. First, the child was invited to complete a short "game" on the laptop, where each of the six target words was introduced three times ("Look, this is a [word]. Do you see the [word]? Click on the [word]."), while the corresponding image was shown on the screen. This was done to familiarize the child with the target words, so that they had some prior knowledge before practicing with the robot. The child was then invited to go sit at the table with the tablet and robot, and play the game of *I spy with my little eye* for 30 rounds as described previously. After completing the interaction with the robot, children were asked once more to sit down at the laptop and complete the English post-test. The total duration of this session was 25–45 minutes, depending on experimental condition — gestures slowed down the training — and on the number of repair rounds needed. The researcher was always present during the session, although he or she was instructed to act busy to avoid having the child turn to them for task-related feedback.

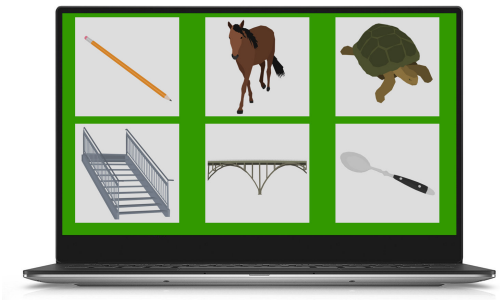

Figure 4: "Game" used to test children's word knowledge.

3.3.4 Delayed Post-Test. Between one and two weeks after the training session with the robot, each child was retrieved from the classroom once more for a delayed post-test. This test was identical to the immediate post-test administered after the child's interaction with the robot, and lasted approximately three minutes.

3.4 Analyses

In line with the preregistration and with the original study, we have conducted a series of ANOVAs with difference scores between the post-tests and pre-test. However, after submitting the preregistration we realized that a single mixed ANOVA would be more optimal, since it reduces the risk of type I errors by minimizing the amount of statistical analyses required. For consistency, we present the results of both analyses. Engagement was annotated by extracting two video clips from each child's interaction with the robot, one from the 4^{th} and one from the 24^{th} round of training. Each clip lasted two minutes and was annotated for two different measures of engagement: task engagement and social engagement with the robot. The ratings were based on a coding scheme that was recently developed[3], which resulted in a score for each type of engagement on a nine-point scale (1–9). Note that engagement is considered as a measure of how actively the child was involved with the robot or the task, not whether this was positive (constructive) or negative (destructive) involvement. The Pearson correlation between task and robot engagement was .60 ($p < .001$).

In comparison to our previous analysis of engagement [8] we aimed to improve robustness by increasing the length of each clip (two minutes rather than five seconds), by rating engagement across two distinct dimensions rather than a single all-encompassing measurement, and by using coding schemes upon which to base these ratings. Instead of distributing an online questionnaire, the ratings were now performed by one of the researchers. To test the reliability of our measures, 50 video clips (taken from 25 different sessions) were annotated by a second rater who did not participate in the original data collection and was not familiar with the specifics of the experimental conditions. The intraclass correlation (ICC) estimates and their 95% confidence intervals were calculated using SPSS version 24 based on a single rater, consistency, two-way random effects model. This resulted in a 95% CI of [.45, .78] for task engagement (considered poor–good, cf. [20]), and a 95% CI of [.55, .83] for robot engagement (moderate–good). Based on this ICC we proceeded with the ratings of a single rater in our analyses.

[3] https://github.com/l2tor/codingscheme

4 RESULTS

4.1 Preregistered Analyses

4.1.1 Learning Outcomes. Figure 5 shows the mean scores on the three tests per condition, indicating a similar increase in vocabulary knowledge over time between conditions. A 3 (experimental condition) × 3 (test time) mixed ANOVA was used to evaluate children's learning outcomes, with scores on the test tasks (0–6) as dependent variable, experimental condition as between-subjects independent variable, and time (pre-test, immediate post-test, and delayed post-test) as within-subjects independent variable. The analysis showed a significant effect of time, $F(2, 182) = 45.70, p < .001, \eta_p^2 = .33$, indicating that children learned L2 vocabulary from their interactions with the robot regardless of condition. Pairwise comparisons using Bonferroni correction show a significant difference between the immediate post-test and the pre-test, $M_{dif} = 1.10, p < .001$, and between the delayed post-test and the pre-test, $M_{dif} = 1.41, p < .001$. However, there was no significant difference between the delayed post-test and the immediate post-test, $M_{dif} = 0.30, p = .09$. There was no main effect of condition, $F(2, 91) = 0.38, p = .68$, and no significant interaction between experimental condition and time, $F(4, 182) = 1.58, p = .18$, indicating that the robot's use of gestures — either repeated or varied — did not affect learning outcomes[4].

4.1.2 Engagement. Figure 6 visualizes task engagement (left) and social engagement with the robot (right), measured at rounds 4 and 24. A clear drop between rounds 4 and 24 can be observed for both types of engagement. Although task engagement levels are similar between conditions, children in the experimental condition without gestures are less engaged with the robot than those in both gesture conditions. To evaluate whether the robot's use of gestures affected children's engagement, we conducted a 3 (experimental condition) × 2 (time) mixed MANOVA with the task and robot engagement ratings as dependent variables, time (round 4 and round 24) as within-subjects independent variable and experimental condition as between-subjects independent variable. This shows a significant effect of time, Wilk's $\Lambda = .30, F(2, 90) = 107.76, p < .001, \eta_p^2 = .71$, indicating a drop in engagement between rounds 4 and 24. This effect was found for task engagement, $F(1, 91) = 132.26, p < .001, \eta_p^2 = .59$, and for robot engagement, $F(1, 91) = 134.79, p < .001, \eta_p^2 = .60$.

The analysis also showed a main effect of experimental condition, Wilk's $\Lambda = .60, F(4, 180) = 13.20, p < .001, \eta_p^2 = .23$, indicating differences in average engagement throughout the interaction. This difference was only significant for robot engagement, $F(2, 91) = 25.9, p < .001, \eta_p^2 = .36$, and not for task engagement, $F(2, 91) = 1.88, p = .16$. A post-hoc analysis using Bonferroni correction showed that average robot engagement was significantly higher in the repeated gestures condition ($M_{dif} = 1.82, p < .001$), as well as in the varied gestures condition ($M_{dif} = 1.93, p < .001$), compared to the condition without gestures. The difference between the varied and repeated gesture conditions was not significant ($M_{dif} = 0.06, p = 1.0$). The interaction between time and condition was not significant, Wilk's $\Lambda = .90, F(4, 180) = 2.32, p = .06$, showing no effect of the robot's use of gestures on the change in engagement over time.

[4]For consistency with the preregistration and the analyses in the original study, we also performed a combination of t-tests and separate ANOVAs on difference scores. The results are identical to the mixed ANOVA approach (a significant effect of time but not condition), with the exception of the difference between the delayed post-test and immediate post-test scores, which now also reached significance.

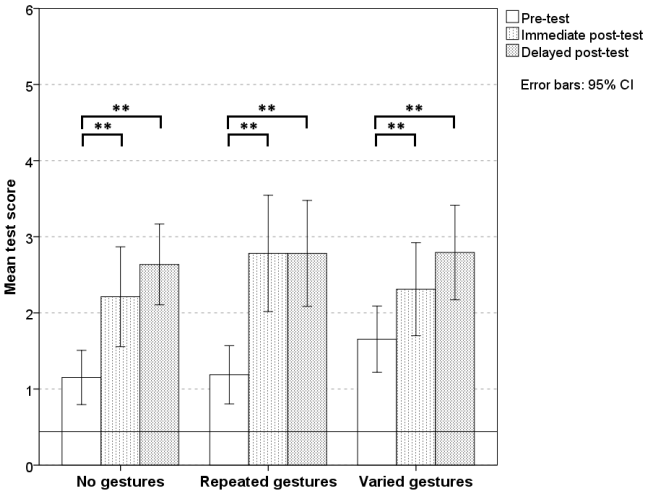

Figure 5: Mean test scores as a function of experimental condition ($p < .001$**). Chance level (horizontal line) was 0.44.**

4.2 Exploratory Analysis of Age

Existing literature indicates that our ability to recognize and understand gestures grows with age [29, 37]. Additionally, we intuitively observed variations in how children of different ages interacted with the robot. Figure 7 shows a linear fit to children's difference scores on the immediate (left) and delayed (right) post-tests, indicating that age affected children's performance, especially in both experimental conditions where the robot used gestures. We ran the same mixed ANOVA with test scores as dependent variable, and time and condition as independent variables, now adding children's age in months at the time of the experiment as a covariate. This showed a significant main effect of age, $F(1, 90) = 19.30, p < .001, \eta_p^2 = .18$. The interaction between age and time was also significant, $F(2, 180) = 10.59, p < .001, \eta_p^2 = .11$, indicating that older children that participated in the study learned significantly more from the interaction than younger children. To further explore whether this effect of age was influenced by the robot's use of gestures, we split our data by experimental condition and ran the same analysis. This showed a significant interaction effect of age and time within the repeated gestures condition, $F(2, 60) = 7.83, p = .001, \eta_p^2 = .21$, and within the varied gestures condition, $F(2, 54) = 7.87, p = .001, \eta_p^2 = .23$, but not within the condition without gestures, $F(2, 62) = 0.74, p = .48$.

To investigate whether age also influenced children's levels of engagement, we ran the previously described mixed MANOVA with both measures of engagement as dependent variables, adding age as a covariate. This showed a main effect of age, Wilk's $\Lambda = .91, F(2, 89) = 4.41, p = .02, \eta_p^2 = .09$. This effect was only significant for task engagement, $F(1, 90) = 5.29, p = .02, \eta_p^2 = .06$, where

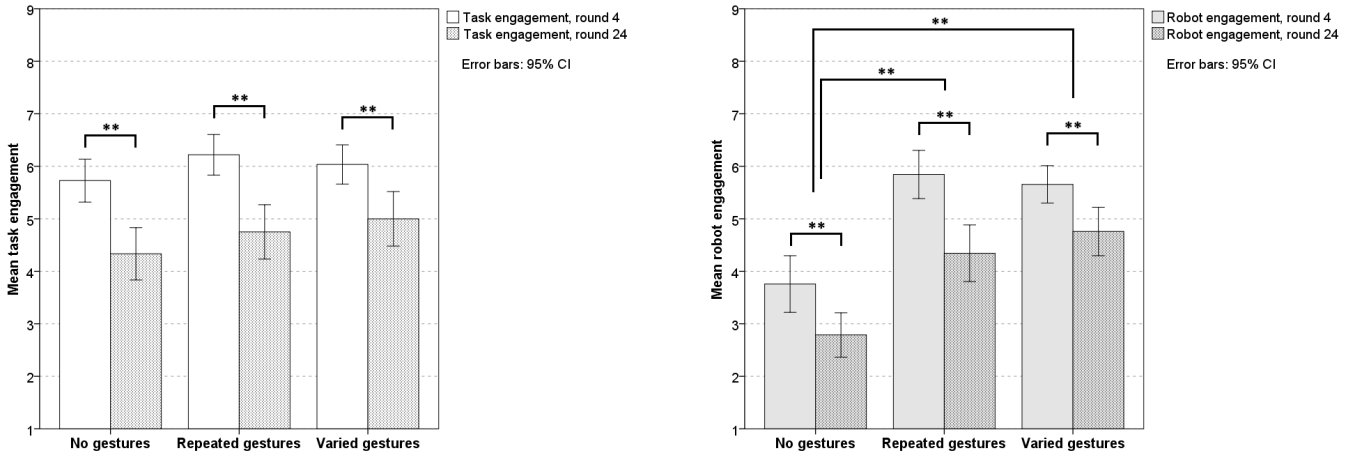

Figure 6: Task (left) and robot (right) engagement ratings for rounds 4 and 24, by condition ($p < .001$).**

the older children in the experiment showed higher task engagement than the younger children. There was no main effect for robot engagement, $F(1, 90) = .002, p = .97$, and no significant interaction effect between age and time, Wilk's $\Lambda = .97, F(2, 89) = 1.18, p = .31$.

5 DISCUSSION

This paper describes a study that investigated the potential benefits of a robot's use of gestures in second language tutoring. We compared between a robot that repeated the same gesture for each concept, one that varied its gesture repertoire, and one that did not use gestures at all, and we measured how this affected children's learning outcomes and engagement with the task and with the robot. The contribution of this work is twofold. First, it is a conceptual replication of a previous study [8] with a shift towards a more diverse set of target words. Our goal was to verify whether our previous findings persist, especially in light of conflicting findings regarding robot-performed gestures in other studies (e.g., [47]). Several steps have been taken to improve the reliability and reproducibility of the study. These include various changes to the measures such as testing each target word multiple times, and the use of a coding scheme for rating children's engagement. Second, despite the assumed importance of variation for educational purposes [24, 31] we did not find any existing research in this direction. Therefore we added an experimental condition where the robot introduced variation by performing different gestures for each concept. Our results show that a single tutoring session with the robot helped children acquire new L2 vocabulary, and retain this knowledge over time. Children on average learned 1.10 new words on the immediate post-test, and 1.41 on the delayed post-test — similar results to those in the original study [8]. This may not seem like a substantial increase, however these were young children and the results were obtained after a single training session of approximately 15 minutes. Other word learning studies with robots have shown similar results [5, 42].

Contrary to the original study we did not find support for our first hypothesis that children would learn and remember more words when the robot used gestures than when the robot did not use gestures. This could be caused by the fact that we introduced more diverse and potentially more complex target words in the current work compared to the animal names in the original study, with perhaps less iconic gestures as a result. Because the overall number of words learned is similar across both studies, we can assume that the English words themselves were not necessarily more difficult to learn. The difference therefore appears to be in the gestures, where children found it harder to understand the gestures in the current study. It would be interesting to further investigate which exact characteristics of the gestures are responsible for these difficulties with their interpretation.

Older children in our study did appear to understand and benefit from the robot's gestures, while younger children did not. Although literature indicates that children learn how to make sense of iconic gestures at a slightly younger age than the age of participants in our study [29, 37], the ability to interpret gestures could be reduced when the interaction involves a robot instead of a human, and when it is mediated by a tablet device. The robot's gestures appear to have a detrimental effect when they are not understood, which may have been caused by distraction, confusion, and the additional cognitive load from attempts to observe and make sense of these gestures. These findings underline the importance of properly designing the robot's gestures. Previous research often included gestures that were designed by the researchers, but in this work we based the design on a dataset with recordings of mostly children performing gestures [9]. The clarity of the robot's gestures was evaluated with 19 judges, and the consistency of the ratings showed that this sample size was sufficient. However, the process of designing gestures could be further improved in two ways. First, it would be better to evaluate the gestures with children from the same age group that participated in our study instead of adults. However, we believe that a task to judge the meaning of gestures is difficult for children this young, so this should perhaps be done in the form of a guessing game. Second, based on the ratings we made several improvements to the gestures, but these were not evaluated. We are confident that these changes resulted in better gestures since they now align more with the original human-performed examples, but in future work we would take a more iterative approach and conduct multiple evaluations.

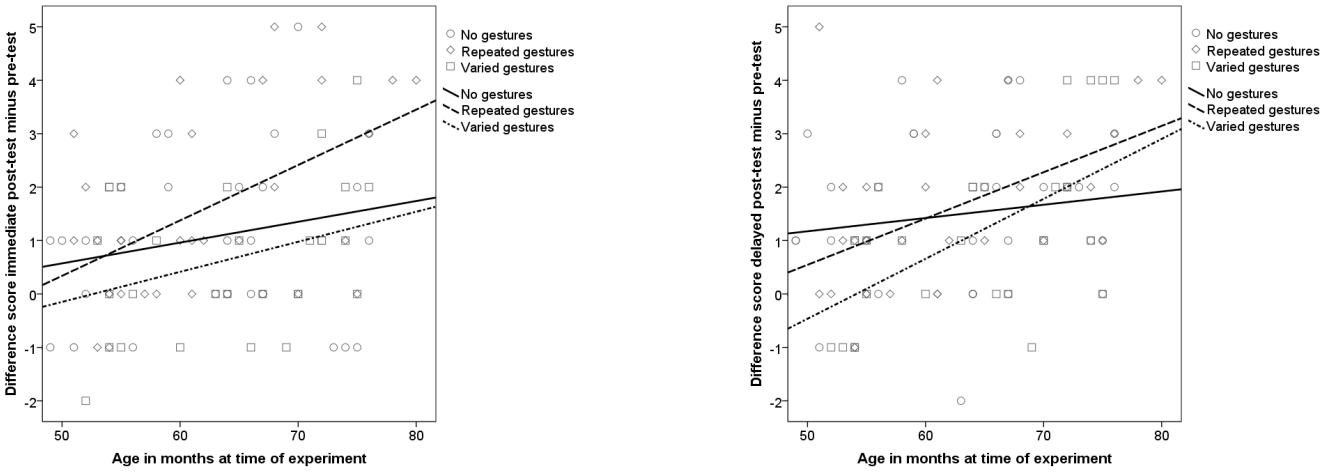

Figure 7: Linear fit to the difference scores on the immediate (left) and delayed (right) post-tests compared to the pre-test per condition, relative to children's age.

Our second hypothesis stated that children would be more engaged with a robot that produces iconic gestures, than with one that does not produce gestures. This hypothesis finds partial support in a higher average robot engagement, however no significant effects on task engagement are found. These findings are consistent with literature on the effects of robot gestures on engagement [7, 8, 11, 35]. We conjecture that the main reason for higher robot engagement is that the robot displayed more bodily movements in the gesture conditions, which can cause the robot to be perceived as more friendly and human-like [2], resulting in a higher level of engagement with the robot as children enjoyed the interaction more. Engagement with the task was influenced by age, however this does not seem to relate to the robot's use of gestures.

By introducing variation in the robot's gestures, and thereby highlighting different features of the object of learning (cf. [24]), we aimed to provide greater support to the learning process compared to using repeated gestures. We also expected this variation in the robot's behavior to further increase children's engagement with the robot (cf. [39]). However, we did not find support for hypotheses H3 and H4 which stated that the robot's use of varied gestures would lead to better learning outcomes and higher levels of engagement than repeated gestures. This does not align with existing findings in literature regarding positive effects of speaker variation [3], nor detrimental effects of image variation [36]. Moreover, with multiple gestures for the same concept it is more difficult to measure what the contribution of each individual gesture was to children's learning outcomes and engagement. We believe more research is needed to further investigate possible differences between variation and repetition of gestures. The current study consisted of a single tutoring session and therefore did not investigate any potential long-term effects that variation in gestures might have. Furthermore, different results could be observed for older children or adults, and the use of varied gestures could have affected other factors that were not measured in the current study, such as perception of the robot (e.g., human-likeness, intelligence, character) or overall enjoyment. With younger participants it remains a challenge to investigate these aspects of a robot's appearance and behavior.

6 CONCLUSION

This paper documents a study that was conducted to investigate whether a robot's use of iconic gestures affects learning outcomes and learners' engagement. Furthermore, a robot that varied its gesture repertoire for a particular concept was compared with one that always repeated the same gesture. The results of the study show that there are advantages to having a robot perform gestures when teaching children L2 vocabulary, in the form of higher engagement and — for the older children in the study — increased learning gain, although no additional benefits were found for varied gestures. Based on existing literature into robot-performed gestures (e.g. [7, 15, 22, 43]) we have reason to believe that our findings generalize to different target groups, educational domains, and robotic platforms, and we imagine that robots in the future will become capable of performing increasingly more human-like motions. The design of the interaction, the gestures, and the study itself are documented in this paper to serve as a basis for future research. We envision two main avenues for future work: (1) the design of the robot's gestures, and how this affects their comprehensibility for different ages, and (2) a further exploration of variation in gestures: does it have different effects on older learners, and does it change the way the robot and the interaction are perceived?

ACKNOWLEDGMENTS

The authors would like to thank Henrike Colijn for her contribution to the annotation of engagement, Martijn Faes for his support with data collection, and the five reviewers for their valuable feedback on the initial manuscript. Finally, we are very grateful to all the schools, teachers, parents, and children that participated in the study.

REFERENCES

[1] Amir Aly and Adriana Tapus. 2013. A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Press, 325–332.

[2] Thibault Asselborn, Wafa Johal, and Pierre Dillenbourg. 2017. Keep on moving! Exploring anthropomorphic effects of motion during idle moments. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 897–902.

[3] Joe Barcroft and Mitchell S Sommers. 2005. Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition* 27, 3 (2005), 387–414.

[4] Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.* 59 (2008), 617–645.

[5] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science Robotics* 3, 21 (2018), eaat5954.

[6] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 708–713.

[7] Paul Bremner, Anthony G Pipe, Chris Melhuish, Mike Fraser, and Sriram Subramanian. 2011. The effects of robot-performed co-verbal gesture on listener behaviour. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 458–465.

[8] 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)*. ACM, 50–58.

[9] Jan de Wit, Bram Willemsen, Mirjam de Haas, Emiel Krahmer, Paul Vogt, Marije Merckens, Reinjet Oostdijk, Chani Savelberg, Sabine Verdult, and Pieter Wolfert. 2019. Playing charades with a robot: Collecting a large dataset of human gestures through HRI. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 634–635.

[10] Marina Fridin. 2014. Kindergarten social assistive robot: First meeting and ethical issues. *Computers in Human Behavior* 30 (2014), 262–272.

[11] Michael J Gielniak and Andrea L Thomaz. 2012. Enhancing interaction through exaggerated motion synthesis. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 375–382.

[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.

[13] Barbara Hanfstingl, Gertraud Benke, and Yuefeng Zhang. 2019. Comparing variation theory with Piaget's theory of cognitive development: more similarities than differences? *Educational Action Research* (2019), 1–16.

[14] Autumn B Hostetter. 2011. When do gestures communicate? A meta-analysis. *Psychological Bulletin* 137, 2 (2011), 297.

[15] Chien-Ming Huang and Bilge Mutlu. 2013. Modeling and evaluating narrative gestures for humanlike robots. In *Robotics: Science and Systems*. 57–64.

[16] 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* 12, 3 (2018), 146–151.

[17] Spencer D Kelly, Sarah M Manning, and Sabrina Rodak. 2008. Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass* 2, 4 (2008), 569–588.

[18] Adam Kendon. 2004. *Gesture: Visible action as utterance*. Cambridge University Press.

[19] Sotaro Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes* 24, 2 (2009), 145–167.

[20] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* 15, 2 (2016), 155–163.

[21] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44, 4 (2012), 978–990.

[22] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37.

[23] 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.

[24] Ference Marton and Shirley Booth. 2013. *Learning and awareness*. Routledge.

[25] Ingrid Masson-Carro, Martijn Goudbeek, and Emiel Krahmer. 2015. Coming of age in gesture: A comparative study of gesturing and pantomiming in older children and adults. *Gespin, Nantes* (2015).

[26] Ingrid Masson-Carro, Martijn Goudbeek, and Emiel Krahmer. 2017. How what we see and what we know influence iconic gesture production. *Journal of Nonverbal Behavior* 41, 4 (2017), 367–394.

[27] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[28] Cornelia Müller. 2014. Gestural modes of representation as techniques of depiction. *Body–language–communication: An international handbook on multimodality in human interaction* 2 (2014), 1687–1702.

[29] Miriam A Novack, Susan Goldin-Meadow, and Amanda L Woodward. 2015. Learning from gesture: How early does it happen? *Cognition* 142 (2015), 138–147.

[30] Gerardo Ortega and Asli Özyürek. 2016. Generalisable patterns of gesture distinguish semantic categories in communication without language. (2016).

[31] Jean Piaget and Margaret Cook. 1952. *The origins of intelligence in children*. Vol. 8. International Universities Press New York.

[32] Emmanuel Pot, Jérôme Monceaux, Rodolphe Gelin, and Bruno Maisonnier. 2009. Choregraphe: a graphical tool for humanoid robot programming. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 46–51.

[33] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.

[34] Kazuki Sekine, Catharine Wood, and Sotaro Kita. 2018. Gestural depiction of motion events in narrative increases symbolic distance with age. *Language, Interaction and Acquisition* 9, 1 (2018), 40–68.

[35] Candace L Sidner, Christopher Lee, Cory D Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166, 1-2 (2005), 140–164.

[36] Mitchell S Sommers and Joe Barcroft. 2013. Effects of referent token variability on L2 vocabulary learning. *Language Learning* 63, 2 (2013), 186–210.

[37] Carmen Stanfield, Rebecca Williamson, and Şeyda Özçalişkan. 2014. How early do children understand gesture–speech combinations with iconic gestures? *Journal of Child Language* 41, 2 (2014), 462–471.

[38] Lauren J Stites and Şeyda Özçalışkan. 2017. Who did what to whom? Children track story referents first in gesture. *Journal of Psycholinguistic Research* 46, 4 (2017), 1019–1032.

[39] Fumihide Tanaka, Aaron Cicourel, and Javier R Movellan. 2007. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences* 104, 46 (2007), 17954–17958.

[40] Marion Tellier. 2008. The effect of gestures on second language memorisation by young children. *Gesture* 8, 2 (2008), 219–235.

[41] Laura Valenzeno, Martha W Alibali, and Roberta Klatzky. 2003. Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology* 28, 2 (2003), 187–204.

[42] Rianne van den Berghe, Josje Verhagen, Ora Oudgenoeg-Paz, Sanne van der Ven, and Paul Leseman. 2019. Social robots for language learning: A review. *Review of Educational Research* 89, 2 (2019), 259–295.

[43] 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.

[44] Karin Van Nispen, Van de Sandt-Koenderman, Lisette Mol, and Emiel Krahmer. 2014. Pantomime strategies: On regularities in how people translate mental representations into the gesture modality. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 36.

[45] Karin van Nispen, W Mieke van de Sandt-Koenderman, and Emiel Krahmer. 2017. Production and comprehension of pantomimes used to depict objects. *Frontiers in Psychology* 8 (2017), 1095.

[46] 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 (2017), 73.

[47] Paul Vogt, Rianne van den Berghe, Mirjam de Haas, Laura Hoffman, Junko Kanero, Ezgi Mamus, Jean-Marc Montanier, Cansu Oranç, Ora Oudgenoeg-Paz, Daniel Hernández García, et al. 2019. Second Language Tutoring using Social Robots: A Large-Scale Study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 497–505.

[48] Elizabeth Wakefield, Miriam A Novack, Eliza L Congdon, Steven Franconeri, and Susan Goldin-Meadow. 2018. Gesture helps learners learn, but not merely by guiding their visual attention. *Developmental Science* 21, 6 (2018), e12664.

[49] Rolf A Zwaan, Alexander Etz, Richard E Lucas, and M Brent Donnellan. 2018. Making replication mainstream. *Behavioral and Brain Sciences* 41 (2018).