International Journal of Social Robotics (2022) 14:827–843
https://doi.org/10.1007/s12369-021-00824-3

# Social Robots for (Second) Language Learning in (Migrant) Primary School Children

**Elly A. Konijn[1] · Brechtje Jansen[1] · Victoria Mondaca Bustos[1] · Veerle L. N. F. Hobbelink[2] · Daniel Preciado Vanegas[1]**

Accepted: 30 August 2021 / Published online: 11 October 2021


**Abstract**
Especially these days, innovation and support from technology to relieve pressure in education is highly urgent. This study tested the potential advantage of a social robot over a tablet in (second) language learning on performance, engagement, and enjoyment. Shortages in primary education call for new technology solutions. Previous studies combined robots with tablets, to compensate for robot's limitations, however, this study applied direct human–robot interaction. Primary school children ($N$ = 63, aged 4–6) participated in a 3-wave field experiment with story-telling exercises, either with a semi-autonomous robot (without tablet, using WOz) or a tablet. Results showed increased learning gains over time when training with a social robot, compared to the tablet. Children who trained with a robot were more engaged in the story-telling task and enjoyed it more. Robot's behavioral style (social or neutral) hardly differed overall, however, seems to vary for high versus low educational abilities. While social robots need sophistication before being implemented in schools, our study shows the potential of social robots as tutors in (second) language learning.

**Keywords** Robot tutor · Robot-child interaction · Second language learning · Field experiment · Longitudinal · Primary school · Migrant children

## 1 Introduction

Primary education in Europe is facing a rising shortage. Budget cuts and shortage of personnel result in growing classrooms whereas they are also facing increasing diversity. The global Covid-pandemic has made the shortage of personnel in primary education even more prominent. As teachers fall sick, under-qualified staff is faced with teaching their classes [11]. The need for innovation in digital education has become greater than ever [4]. Qualified teachers are scarce, in particular for special need students, and nearly half of schools report a shortage of teachers for these students [30]. The shortage is most prominent in schools with a complex student population where some students require more attention than others [29]. Primary school classes today exist of children with different educational levels, cultural backgrounds and socioeconomic status. As of today, almost one-third of children starting primary school in the Netherlands has a migration background [29]. Children with a migration background start primary school with a disadvantage, as their educational achievement falls behind even at the age of three [40, 55]. These unequal levels of knowledge further complicate compiling a curriculum for preschool teachers relevant to all children in their class [37]. This is especially relevant as children profit most from education adjusted to their level of knowledge [61].

To cope with the growing shortage of staff and increased diversity, technologies such as social robots and tablets could offer opportunities for improvement. Tablets are accessible tools and currently, primary schools have started using them for tailored learning tasks. Learning with a tablet seems to affect learning outcomes positively, however, a number of difficulties have emerged for implementing tablets as learning tools [27, 58]. Challenges with using tablets for education purposes are, for instance, the device's distracting nature, the abundance of applications, technical issues

✉ Elly A. Konijn
elly.konijn@vu.nl

[1] Department of Communication Science, Media Psychology Program, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2] Department of Psychobiology, University of Amsterdam, Amsterdam, The Netherlands

and expertise needed for their use and implementation in teaching programs [58]. The design of a tablet is reported to be problematic with many easy-accessible and distractive apps, which raises pedagogical questions around duration of engagement with the tablet and what applications are accessible to children during this time [48]. Thus, using tablets for educational purposes seems to come with many challenges.

An alternative option worth exploring is a social robot. A great advantage of social robots over tablets is its embodiment and anthropomorphism [8, 9]. People easily ascribe human features to a social robot rather than to a tablet [57]. Humanoid robots are more likely to be perceived as friends with whom people enjoy interacting more than with a virtual agent in a tablet application [51]. The tablet's lack of embodiment also causes people to trust it less than an embodied robot and people follow instructions from a robot more closely than similar instructions provided by a tablet with the same voice [43]. These findings support the importance of both the physical appearance and the social capabilities of interactive technology, and embodied robots that resemble human beings both physically and behaviorally seem to have the advantage in terms of acceptance, trustworthiness and likeability. Importantly, this increased likeability and acceptance of social robots might give them an additional edge as tutors and educational aids, compared with other forms of non-social technologies, such as tablets. Therefore, it might be worthwhile to explore using social robots as a replacement for tablets, as they do not carry other non-educational applications, which in turn might reduce distraction.

## 1.1 Theoretical Framework

Social robots are shown to be effective tutors, enhancing concentration and academic performance, and are used in language development in young children [7]. Social robots have the potential to enable one-on-one tutoring and tailored exercises for which teachers currently do not have the time nor capacity, whereas one-on-one tutoring is most effective [12]. In addition, language learning with a robot tutor can be experienced as less intimidating in comparison to a peer or teacher [23]. In contrast to their human counterparts, social robot tutors will never get tired or annoyed, and are unbiased towards students. Training can be adjusted to the individual's knowledge level, which is valuable in (second) language learning [49, 61]. However, studies also show mixed results for robot tutors, in particular in the area of second language learning, and have methodological problems like small sample sizes, lack of adequate control conditions in the experiments, and not accounting for potential confounding factors that may have affected the participant's performance [8]. Clarity around the true effect of robot tutors, and particularly social robot tutors is needed in order for them to be implemented, especially when working with a young target audience.

### 1.1.1 Embodiment, Perceived Humanness and Interaction

Importantly, the embodiment of social robots is advantageous as it enables the possibility of natural interaction, which is beneficial for (second) language learning [8]. Natural interaction is defined as communicating through gestures, expressions, movements and manipulating physical objects by interacting with the real world [59]. At a young age, children learn language through conversation with their parents, storybook reading, and other use of language in their environment [10, 14, 22]. Through embodiment social robots can provide natural interaction, which stimulates language learning in children. Natural interaction is further stimulated through a robot's human-like features (e.g., eyes, mouth). An embodied system such as a humanoid social robot is more stimulating for oral dialogue, and therefore language development, than devices that lack these human-like features [8, 53].

This perceived humanness, or how human-like an object or agent is evaluated, gives a social robot advantages as a tutor. Human-like features in a system elicits it being anthropomorphized [18]. Anthropomorphic robots are commonly used in (second) language learning, and support attribution of the robot as real conversational partner, as well as boosting student engagement (Randall 53). A recent study in second language learning found that children who perceived the robot as more human, knew more words after tutoring [9]. In interacting with humanoids, people can even feel an affective bond between themselves and the robot [33]. This perceived humanness creates potential for social robots to take up the role of teacher, tutor or peer for learning tasks.

### 1.1.2 Enjoyment and Engagement

Additionally, prior studies showed that enjoyment and engagement during learning exercises can increase learning gains [24, 35] further confirmed by a meta-analysis [7]. Children who engaged more in story-telling with a robot also showed higher learning outcomes [49, 56]. The use of feedback can be of help to increase engagement in a (second) language learning task [1, 26]. Furthermore, language learning with a social robot boosts enjoyment, which facilitates a better learning experience, causing the students to immerse themselves in the learning process [2]. Therefore, enjoyment and engagement are important to include when studying (second) language learning. However, van den Berghe et al. [8] point out that high engagement might also be due to the novelty effect of interacting with the robot, implying that the beneficial effects of the robot on engagement and enjoyment may diminish as the robot tutor becomes more familiar and

commonplace. Thus, while engagement and enjoyment possibly influence the learning process positively, confounding effects need to be accounted for.

#### 1.1.3 Robot Behavior and Education

In education, a teacher's behavior or communication style is an important differential factor in student achievement gains [62]. Teachers can communicate in a neutral or more personalized and socially supportive way, and for human teachers the latter is usually more successful [21, 39, 65]. Studies with robot tutors in education thus far applied two types of robots' behaviors: including social behaviors like gestures, or acting just neutral and strictly task-oriented. However, results are generally mixed [7]. A number of studies report different effects, or lack thereof, on learning outcomes when robots express different gestures, gazes and behavioral expressions [15, 31, 32, 36, 42, 63]. There is a fine line between the robot being social enough to sustain children's interest and the robot distracting or even intimidating children by being too social [8].

In varying a robot's social style, the specific characteristics of the robot's learner may make a difference. Individual differences in attentional control (e.g., ADHD, [3]) can complicate the learning process and cause a child to be easily distracted. Varying the communication style (social vs. neutral) in a study using an autonomous embodied robot rehearsing multiplication tables impacted children with individual differences in math ability (below vs. above average) differently [32]. Particularly, children whose skills were below-average in math ability did not benefit from the robot's social communication style and performed better with a neutral robot. As opposed to a high-concentration task like math, language learning requires social communication as an integral part to support natural interaction [28, 59]. Varying a robot's social style might therefore have different effects in (second) language learning.

Thus far, however, research comparing a social robot to a tablet has not yet found significant differences in learning outcomes for (second) language learning [8, 37, 60]. We argue that an important limitation in those studies was that an accompanying tablet was used alongside the robot for the interaction. The tablet was needed as input device because Automatic Speech Recognition (ASR) for the ability to accurately process spoken language in robots is not yet sufficiently developed. Problematic is that tablets distract the focus of attention from the robot, or even the task altogether and lead the child to primarily focus on the tablet [58, 60]. This limits the natural child-robot interaction and reduces the value of the robot as a physical entity [7, 8, 13, 32]. A mediating device, such as a tablet, may undermine the advantage of the robot being a physical conversational partner in direct interaction with the pupil [34]. Particularly in language education, where 'face-to-face' conversation is key to natural interaction, this might be problematic. As stated earlier, natural interaction is important for language learning and would lead to better understanding of the effect of robot tutors. Therefore, the current study crucially differs from previous studies in studying direct interaction with a robot, that is, without using a tablet in between to allow for direct robot-child interaction.

In sum, the aim of this study is to test the effectiveness of a social robot without an additional tablet for (second) language learning. We focused on children aged 4–6 years, learning Dutch as first or second language with a socially or neutrally behaving robot compared to a tablet. Most participants have (parents from) a (non-western) migration background and just entered primary school in the Netherlands. In a 3-wave (T0-T2) field experiment, the robot/tablet read an interactive story with the children. Interactive storytelling is a natural way for young children to learn new vocabulary and has proven to be effective in a situation where parents or teachers read a story [10, 44, 45]. T0 served as a baseline and in between T1 and T2, children were read stories 3 times including exercises (E1-E3). Learning outcomes, engagement and enjoyment of language learning exercises with a (social or neutral) robot were compared to the same exercises with a tablet. We hypothesized that a humanoid robot will lead to higher learning outcomes (H1) and higher enjoyment and engagement (H2) in language learning exercises than a tablet. Perceived humanness was included in this study as potential mediator (H3) in between embodiment of the device and learning outcomes, engagement and enjoyment. As an additional exploration, the effect of behavioral style of the robot on learning outcomes (RQ1) was analyzed.

The following section provides a detailed description of the experimental design, protocols, experimental conditions and collected measures of learning outcomes and subjective assessments of engagement, enjoyment and perceived humanity. Thereafter, we describe the results of our hypotheses testing analyses, enhanced with exploratory analyses to gain further insights regarding the factors that influence learning outcomes in this context. At last, results are discussed in light of existing literature, discussing also the limitations and implications of the present study.

## 2 Material and Methods

The following sub-sections describe the details of the evaluated sample, including relevant demographic and experimental variables and how they were measured and calculated, as well as the technical and logistic details of the study protocol.

## 2.1 Participants and Design

All participants were Dutch primary-school children aged between 4 and 6 ($N = 63$, $M_{age} = 5.45$ years, $SD = 0.68$, 39 boys) who participated voluntarily after having received active consent from their parents. Teachers of all four classes participating in this experiment provided lists with demographic and other variables that were important to control for (i.e., birth date, gender, parents' background, whether children were raised monolingual or bilingual). 84.1% of children has parents with a migration background and 57.1% of children is bi- or multilingual.

Moreover, teachers gave a score between 1 and 4 for both receptive and productive vocabulary level of each child. From the latter, a new variable was calculated based on the mean outcomes of these two variables, named language level ($M_{LangProf} = 2.40$, $SD = 0.60$, with a score of 2 considered as average performance). Hence, we controlled for language level (receptive vocabulary, productive vocabulary, and the mean of these two) between all conditions at the start of the experiment. Furthermore, the baseline measurement (T0) assessed whether a child had an initial knowledge level beyond or below the mean T0. Participants were recruited in four classes on two primary schools in the Netherlands.

Within each class, children were randomly assigned in a mixed factorial design with device (tablet, $N = 20$, robot; social, $N = 23$, neutral, $N = 20$) as a between factor and learning outcomes (T0: baseline, T1: immediate and T2: delayed post-test scores) as a within factor. During the experiment, a tablet or robot read three interactive story exercises with the individual participants (E1: first, E2: second, and E3: third exercise session). The baseline measurement (T0) was taken 1 week before the first tutoring exercise (E1) to control for prior knowledge related to their individual linguistic skill. The immediate test (T1) was taken directly after E3 and the delayed post-test (T2) was taken 2 weeks after the final story-telling session measuring how many words lingered. Over the course of three weeks, each child had three tutoring sessions, once a week, with either the robot or a tablet in similar ways (Fig. 1). Testing was done either for differences between devices (i.e., robot vs. tablet) or differences between robot's behavioral style (i.e., social vs. neutral). Learning outcomes, engagement and enjoyment were the dependent variables. Perceived humanness was included as a potential mediator.

## 2.2 Materials and Procedures

The following sections provide details regarding the technical specifications and procedures employed for the different experimental conditions, both in terms of device (robot/tablet) and device behavior (social, neutral, or tablet).

### 2.2.1 Robot versus Tablet Conditions

We used a SoftBank NAO robot, 58 cm tall humanoid [25] with Choregraphe 2.4.1. software, which is often used in empirical research and practice [7]. In the tablet condition, an iPad 2 was used with a custom-made web application that was running on Safari. Both conditions were made as comparable as possible. Both the robot condition and the tablet condition made use of the same voice, i.e., the standard voice of robot NAO with speed adjusted in Choregraphe to 80%. The feedback options were also the same for both conditions, as were the images used during the learning tasks. However, responses were adjusted to the relevant device, e.g., the tablet asked to click on the right picture and the robot asked to point out the right picture.

To minimize possible novelty effects, children had an extensive introduction to the robot [8]. This allowed children to get comfortable with the robot. Such an extensive introduction was not deemed necessary for the tablet condition, considering that tablets, smart-phones and comparable devices are frequently used by children this age-group in most households and educational settings in The Netherlands. Thus, we did not expect the tablet to elicit particularly strong novelty effects, especially in contrast to the robots. Experimental exercises with the robot/tablet took place in a separate small room at school. Children were invited one by one from the classroom during the day to have a one-on-one tutoring session. An experimenter was always present in the room, placed behind the child so as not to distract it (Fig. 2).

### 2.2.2 Robot Behaving Socially or Neutral

Both the social and neutral robot were task-oriented and not programmed in a personalized way (i.e., not adapting to the participant's skills). Given the mixed results regarding robots' behavioral style [7], this study programmed the robot's social behavior without gestures or movements during the task and focused on social behavior in dialogue. The social robot welcomed the participant by introducing itself and asking for the participant's name, which it used a number of times throughout. The social robot also reacted to other social cues from the participant such as limited gestures (waving), looking at the participant, but not following eye movements. After the entire exercise, the social robot thanked the child for reading a story together and said "Goodbye". The neutral robot was strictly task-oriented and did not use any social cues. Three feedback options were programmed in the robot ("correct", "false", "try again") as a response to the child's answer (see 'target words').

Automatic Speech Recognition of today's robots is not yet sufficiently developed, specifically not for young children [7]. Therefore, a Wizard of Oz (WOz) technique was complemented to the ASR to minimize errors and simplify

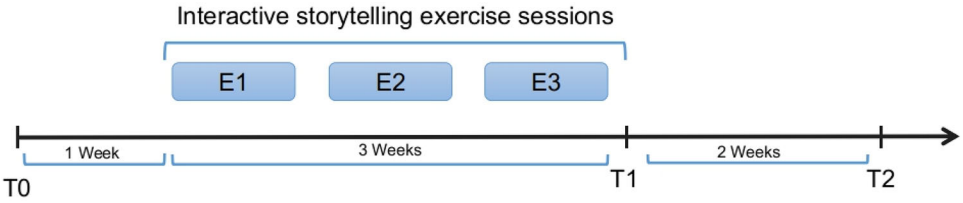


**Fig. 1** Visualization of experimental protocol. *Note.* Interactive storytelling sessions (E1, E2, E3) were held once a week. Learning outcomes were measured at Baseline (T0), Immediately after last exercise session (T1), and 2 weeks after the last exercise session (T2)

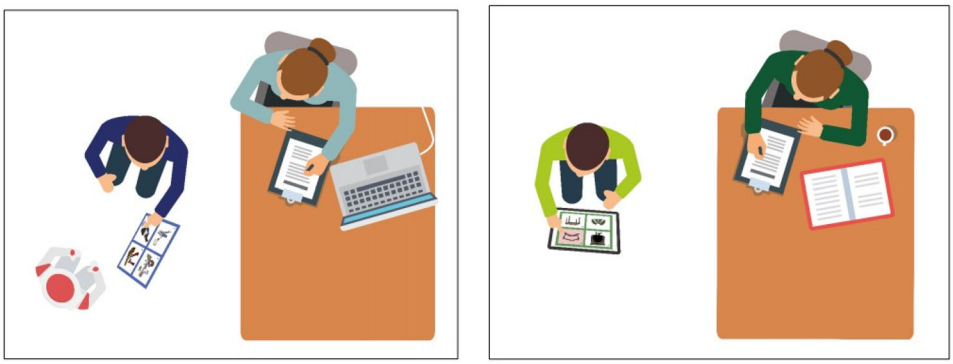

**Fig. 2** Visualization of experimental setting for robot (L) and tablet (R) condition. *Note.* Left: Schematic overview of the setting during exercises in the robot condition. Right: Schematic overview of the setting during exercises in the tablet condition

programming. To ensure that the use of the WOz technique was as inconspicuous as possible, an experimenter was always present in the room behind the child to observe the interaction on both robot and tablet conditions (Fig. 2). This allowed us to avoid discrepancies in the session environment between robot and tablet conditions that might influence the child performance or perception of the robot/tablet, such as realizing that the device is controlled by the present experimenter.

#### 2.2.3 Tablet Condition

The procedure for the tablet condition was largely the same as the robot condition. However, the tablet could not introduce itself, but asked for the child's name. Furthermore, the tablet asked the participant to click on the correct picture on the screen, instead of pointing at it on a physical sheet (Fig. 2). In order to prevent possible disappointment of the children in the tablet condition, because they did not get to play with the robot, these children got the opportunity to read a short story with the robot and ask the robot questions afterwards.

### 2.3 Word Learning through Storytelling

The following subsections describe the details of the interactive story exercise sessions, and how are they implemented in the different experimental conditions defined for the study.

#### 2.3.1 Stories

Target words were explicitly taught whilst the social robot or tablet was telling a story. Three stories, one for each tutoring session, were selected based on the children's age (4–6) and average language level after consulting pedagogues and experts in language development of children in this age category. The three original stories were adjusted to fit the target words. The average duration of the stories, including all three feedback options for every target word, was 10 min and 31 s.

#### 2.3.2 Target Words

Target words and the receptive vocabulary task (i.e., choosing the matching image for each target word) are based on the Peabody Picture Vocabulary Test Fourth Edition [19]. Prior to the experiment, the robot's speech technology and appropriate difficulty level of target words for the age group was pilot tested ($N = 5$). Based on the outcomes of this pilot some words were adjusted.

At fixed points during the storytelling, the participants were asked to match images with the target words. In both robot conditions, these images were printed on a plasticized colored sheet standing on the left side of the child (Fig. 2). Each sheet presented four images from which the child had to choose the correct one. The other three images were related to the target word (i.e., in the same category) but not correct. For example, for the target word'ambulance', three other images of vehicles (e.g., a van, a camper, and a fire truck) were presented alongside the image of the ambulance. In total there were 20 target words, and one word that was used during the

explanation of the exercise. Target words that were covered in previous exercises were incorporated in the text of the following story-telling exercise. Thus, children repeated words from previous sessions in the new session in order for them to remember the target words better. These words were also included in the productive vocabulary test the participants took after each exercise.

Each target word was used in the context of the story first, after which the robot or tablet would repeat the word and ask to point to or click on the matching image. When the child chose the correct image, the device complimented their effort and described the target word as shown on the picture, and asked the child to repeat the word. If the first attempt was not correct, the device would ask to try again and repeat the question. If it picked the right picture the second time, the device complimented the child in the same way as was described above. If not, the robot/tablet would describe the right picture and ask to repeat the target word.

## 2.4 Measures

The following sections describes the different measures employed in the study.

### 2.4.1 Learning Performance

Learning outcomes were measured by the vocabulary tests with the 20 target words. These tests were based on CELF-tests [50], focused at productive vocabulary and measured if children's productive vocabulary had been expanded after the language learning exercises. When answers were incorrect, the child was encouraged to further try 2 times (see 'target words'). Two points were scored if the child got the word correct at first attempt, one point if correct at second attempt, and zero points if it failed.

At the first storytelling exercise (E1), the task contained seven target words, thus resulting in a scoring range between 0 and 14 points. Each session tested new words, as well as repeating the words exercised during the previous session(s). The second storytelling exercise (E2) thus contained six additional target words and could score between 0 and 26 points. The last exercise (E3) added seven target words, thus a score range between 0 and 40 points.

Thus, for the baseline test (T0), the immediate post-test (T1) and the delayed post-test (T2), the minimum score was 0 points (in case they knew none of the twenty target words) and the maximum score was 40 points (if they knew all twenty target words). At T0, how many of the 20 target words the children already knew beforehand was recorded. The delayed post-test (T2) measured how many target words practiced during the exercises lingered over time. After each exercise (E1, E2, and E3), an immediate post-test was conducted for the six or seven target words of the specific week and target words of the prior weeks. Thus, T1 ('immediate post-test') was calculated by the sum of points for the target words at the last storytelling exercise, which included all target words used in the previous exercises as well as the new ones (i.e., E1-E3).

In sum, the dependent variable 'learning outcome' was measured through the scores (0–40 points) on three different time points; the baseline test (T0), immediate post-test (T1), and delayed post-test (T2). In repeated measures ANOVA analyses, these time-based scores were included as within variables. To pinpoint differences, $t$-tests or Wilcoxon tests were used. All cases of significance were determined at the $\alpha = 0.05$ level.

### 2.4.2 Enjoyment, Engagement, and Perceived Humanness

The dependent variables enjoyment and engagement with the interactive storytelling exercise were measured at each storytelling session (i.e., at E1, E2, and E3), obtained through self-report and observation. Perceived humanness of the device was measured only at the last tutoring session through a questionnaire. This way, children had time to get used to the device and form an impression of its 'humanness'.

Engagement and enjoyment were partially determined through an observation scheme the researchers filled out during the exercise and partially through questionnaires that included self-report items. To ensure inter-rater reliability, the experimenters both completed the observation forms for 35% of the cases in the first week. Krippendorff's alpha was $>0.8$ for all items, except for one item on the enjoyment scale ("fun", $a = 0.71$). This was due to the lack of deviation among the given scores (i.e., only scores 1, 2, or 3 were given on this item, not 4). This causes an error in the Krippendorff's alpha calculation, therefore an agreement percentage for "fun" was determined. In 87.5% of the cases, the experimenters agreed on the score of this item. Furthermore, experimenters discussed the specific definition of this item before continuing data collection.

Engagement was originally scored on a 12-item observation scheme during E1, E2 and E3, but to improve consistency and reliability of the scheme, one item was removed from the variable ($N = 11$, Cronbach's $\alpha_{E1} = 0.82$, $\alpha_{E2} = 0.83$, $\alpha_{E3} = 0.80$). The items in this scale were based on prior tasks related to child-robot interaction observation schemes [6, 38, 52], literature and experiences during the pilot tests. Observational items were, for example, if the child was focused on the robot and concentrated on the task. The observations were measured on 4-point Likert scales (1 = not at all, 4 = very much).

Enjoyment was scored through a 7-item scheme during E1, E2 and E3, with both observational and self-report items ($N = 7$, Cronbach's $\alpha_{E1} = 0.76$, $\alpha_{E2} = 0.81$, $\alpha_{E3} = 0.78$). These items were based on literature and existing measure-

ment scales for enjoyment [17, 24, 46, 52]. Enjoyment of children is difficult to measure, because of the so-called yes-bias, referring to the tendency to answer "yes" to closed questions [47]. In trying to bypass young children's yes-bias and limit socially desirable answers, we decided to measure enjoyment not only by asking the children how they liked the stories, exercises and interactions, but also by observing them during the interaction. Thus, enjoyment was measured by 4 items in the observation scheme and 3 self-report items asked during the evaluation of the exercise. Furthermore, these self-report questions were asked as two-part questions, in which the child first had to indicate whether they liked it or not. Thereafter, the child could give an indication of how much they (dis)liked it ("Did you like the story or not?", then, "How much?"). Both the observations and the evaluation questions were measured on 4-point Likert scales (1 = not at all,4 = very much). Four-point scales were advised for this age group by experts (e.g., preschool teachers and pedagogues).

Measuring perceived humanness was based on prior literature and scales measuring anthropomorphism and perceived humanness [5], Duffy [20], [33]. It was originally scored on a 7-item scheme with both observational and self-report items, but to improve reliability, one item was removed ($N = 6$, Cronbach's $\alpha = 0.81$). Perceived humanness was measured through an oral questionnaire, tailored for this age group, only administered at the last storytelling exercise (E3). The questionnaire measured to what extent the children perceived the device they used during the storybook reading exercise as humanlike, on 4-point Likert-type scales (1 = not at all; 4 = very much).

The observation schemes and questionnaires (Translated from Dutch) employed to collect these measures are available in the online Supplementary materials (OSF Repository).

## 3 Results

We describe our findings in terms of preliminary analyses as an initial exploration and processing of raw data, followed by the statistical hypotheses testing and a finals section describing exploratory analysis further investigating possible interactions between the different evaluated variables and their effect on learning outcomes and subjective measures of engagement, enjoyment and perceived humanness.

### 3.1 Preliminary Analysis

Preliminary analyses consisted of the identification and removal of outliers, a characterization of the studies sample and an preliminary analysis of the correlations between variables.

#### 3.1.1 Outliers

Initially, 68 children started this experiment, but 1 child did not complete the three weeks due to absence. Outliers were identified based on their learning outcomes at any of the measurement time-points. To avoid random data elimination, rejection criteria were set in place. There was omission if a participant scored in the lowest or highest 5th percentile for 3 categories: baseline T0, immediate post-test T1, or delayed post-test T2 (2 weeks after last session). For T0, it meant rejection of values of 3 or lower and values of 33 or higher. For T1, rejection of values of 10 or lower and of 39 and higher. Lastly, for T2, rejection of values of 7 or lower and of 39 and higher. Based on these criteria, 4 children were omitted from the data due to extremely low or high scores at any of the time-points where learning outcome were measured, which were not representative to the task. For 3 participants all criteria were applicable with 2 participants scoring in the lowest 5th percentile, and one in the highest 5th. One other child was omitted who scored higher than the criterion in the baseline measurement (baseline score = 4), but proceeded to score lower than the baseline in the immediate and delayed post-test (immediate test score = 2, delayed test score = 2).

Further underpinning this decision, some children were still really young and their language level may simply not be sufficient to comprehend the task and stories. The 3 children who had the extremely low scores were aged only 4.5 years (averaged), whereas the child who scored extremely high was aged 6.3 years. For the latter, the task was possibly too easy and the child had little to gain from the task at this stage of language development, further confirmed by the teachers' scoring of language level. In contrast, the 3 outlier children at the low end, each had below average receptive and productive language levels (score 1).

#### 3.1.2 Descriptive Statistics and Correlations

In all, 63 children participated, aged between 4 and 6 (mean age of 5.45 years ($M_{\text{age in months}} = 65.44$, $SD = 8.17$); 39 boys (61.9%); 24 girls (38.1%)). A full overview of means and standard deviations can be found in Table S1. Most 53, children had parents from a non-western migration background (84.1%) and 10 had parents from a Dutch background (15.9%). There were 36 bilingual or multilingual children (57.1%) and 27 monolingual children (42.9%).

To check whether the participants were randomly assigned to the experimental conditions, a Kruskal–Wallis rank sample test was conducted for different variables (Table S2). No significant differences were found for any of the demographic variables and initial language level. Moreover, there were no significant differences in baseline scores between the tablet and robot conditions (both behaviors included, $t(61) = 0.82$,

$p = 0.42$). Thus, the participants were randomly assigned to the conditions.

The mean score of teacher's evaluation of each child's receptive vocabulary level was 2.6 ($SD = 0.6$) for productive vocabulary 2.2 ($SD = 0.7$), with a combined mean of 2.4 ($SD = 0.6$) for language level. Baseline (T0), mean initial language level was 16.4 ($SD = 8.7$). The first exercise (E1) resulted in seven known target words ($M = 7.03$, $SD = 4.8$), the second exercise (E2) resulted in $M = 15.1$ words ($SD = 6.4$). The immediate post-test (T1, obtained at E3) had a mean of 27.1 words ($SD = 9.0$). Delayed post-test scores (T2) were $M = 26.1$ ($SD = 9.6$).

Engagement over all three measurement times (E1, E2, E3) was $M = 3.0$ ($SD = 0.4$) and enjoyment scored $M = 3.0$ ($SD = 0.5$). Perceived humanness overall was $M = 2.8$ ($SD = 0.7$) and children liked the stories with $M = 3.4$ ($SD = 0.6$).

Correlations between all demographic variables, initial learning level, perceived humanness and the dependent variables in this study are reported in Table 1. Language level (scored by the teachers) and T0 both indicate the children's initial language level and correlate significantly ($r(63) = 0.51, p < 0.001$). Conversely, language level did not appear to correlate significantly with neither the migration background of the child's parents ($r = -0.17, p > 0.05$) nor with whether they were mono-, bi-, or multilingual ($r = -0.07, p > 0.05$).

Furthermore, there is a significant, but not strong, negative correlation between background and baseline (T0) scores ($r(63) = -0.30$, $p = 0.02$) indicating a consistently lower score for children from migrant backgrounds. Perceived humanness correlates strongly with enjoyment ($r(63) = 0.71$, $p < 0.001$) and engagement and enjoyment also significantly correlate ($r(63) = 0.52$, $p < 0.001$).

## 3.2 Testing Hypotheses

### 3.2.1 Testing Hypothesis 1: Effect of Embodiment on Learning Outcomes

To test H1, children who read stories with the robot remember more target words than children who used a tablet, a 2 (device) × 3 (time of testing: T0-T2) repeated measures mixed ANOVA was applied. Greenhouse–Geisser correction was applied because the assumption of sphericity was violated ($W = 0.67$, $p < 0.001$, GGe = 0.75). Levene's test revealed the assumption of equality of variances was not violated. Not all groups were normally distributed, but the repeated measures ANOVA is robust for this violation [64].

The effect of time on learning outcomes was significant ($F(2, 122) = 289.30$, $p < 0.001$, $\eta p^2 = 0.23$), indicating the learning task was successful (Fig. 3). There is no significant main effect of device ($F(2, 61) = 2.81$, $p = 0.14$, $\eta p^2 = 0.03$), however, there is a significant interaction effect between time of testing and device ($F(2, 122) = 4.05$, $p = 0.03$, $\eta p^2 = 0.004$). Children learned more with the robot over time, from the baseline to the delayed post-test, compared to the tablet (see Suppl. Mats., 'Extended analyses H1' for an in-depth analysis), supporting H1. To test for differences between the robot behavior types (RQ1), a 2 (robot behaving socially versus neutrally) × 3 (time of testing: T0-T2) mixed ANOVA was conducted, applying a Greenhouse–Geisser correction. Results yielded no significant interaction effect of time and device ($F(2, 82) = 0.15$, $p = 0.79$, $\eta p^2 < 0.001$), but the effect over time was again significant ($F(2, 82) = 204.67$, $p < 0.001$, $\eta p^2 = 0.27$). Thus, answering RQ1, children did not learn significantly more with either the socially or neutrally behaving robot (see Suppl. Mats., Figure S1).

### 3.2.2 Testing Hypothesis 2: Effect of Embodiment on Engagement and Enjoyment

To test if engagement and enjoyment were higher with a robot than a tablet (H2), a MANOVA was conducted due to the relatively strong correlation between these two dependent variables ($r(63) = 0.52$). For this analysis, we included device (robot, tablet) as between-subjects factor and time (T0, T1, T2) as within-subjects factor, with engagement and enjoyment of the device during interactive storytelling at E1, E2 and E3 as dependent variables. The assumptions of sphericity and homogeneity of variance were not violated and data were not normally distributed. Results showed a significant main effect of device ($F(1, 61) = 36.15$, $p = < 0.001$, $\eta^2 = 0.37$), no interaction, showing that engagement and enjoyment were higher with the robot at each measurement time (Figs. 4 and 5). Furthermore, engagement was a significant predictor of immediate learning outcomes (see Suppl. Mats., 'Extended analyses H2').

To test for differences between the two behavior types of the robot (i.e., social versus neutral), a mixed MANOVA with the independent variable 'robot types' was conducted and dependent variables engagement and enjoyment, repeated for each measurement time-point. Neither the assumption of sphericity nor the equality of variances was violated and data were not normally distributed. Results did not show significant main effects of time or device, nor an interaction effect between time and device. This means children did not engage more in exercising with a social robot than with a neutral robot, and did not enjoy exercising with one more than the other.

### 3.2.3 Testing Hypothesis 3: Effect of Embodiment on Perceived Humanness

To test H3, predicting higher perceived humanness for the robot (social and neutral) than tablet, an ANOVA was conducted with type of robot (social, neutral, tablet) as

**Table 1** Correlations of demographic variables, learning outcomes, perceived humanness, enjoyment and engagement

| | Age | Male | Class | Background | Bilingual | Rec | Prod | Language level | T0 | T1 | T2 | PH | Enjoyment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Age | 1 | | | | | | | | | | | | |
| Male | − .13 | 1 | | | | | | | | | | | |
| Class | − .04 | − .01 | 1 | | | | | | | | | | |
| Background | .04 | .07 | − .12 | 1 | | | | | | | | | |
| Bilingual | .08 | .09 | .10 | .41*** | 1 | | | | | | | | |
| Receptive voc | .31* | .01 | − .08 | − .12 | .01 | 1 | | | | | | | |
| Productive voc | .15 | .04 | .06 | − .15 | − .08 | .63*** | 1 | | | | | | |
| Language level | .24 | − .00 | .03 | − .17 | − .06 | .83*** | 95 | 1 | | | | | |
| Baseline (T0) | .37* | − .41*** | .29* | − .30* | − .23 | .39*** | .47*** | .51*** | 1 | | | | |
| Immediate post test (T1) | .27* | − .47*** | .39** | − .38** | − .23 | .37** | .43*** | .91*** | .97*** | 1 | | | |
| Delayed post test (T1) | .29* | − .47*** | .39*** | − .38** | − .23 | .37** | .37** | .43*** | .91*** | .97*** | 1 | | |
| Perceived humanness | .03 | .02 | − .09 | − .09 | .03 | .02 | .03 | − .01 | − .14 | .07 | − .07 | 1 | |
| Enjoyment | − .03 | .04 | .02 | − .06 | − .08 | − .07 | .02 | − .04 | − .06 | .07 | .71*** | 1 | .52** |
| Engagement | .26* | − .25* | .19 | − .07 | .04 | .07 | .21 | .19 | .38** | .39** | .44*** | .37** | |

Rec. = receptive vocabulary; Prod. = productive vocabulary; T0 = baseline test; T1 = immediate post-test; T2 = delayed post-test; PH = perceived humanness
$N = 63$
***Correlation is significant at the 0.001 level (2-tailed);
**Correlation is significant at the 0.01 level (2-tailed);
*Correlation is significant at the 0.05 level (2-tailed *a*). Language level is the mean score of receptive and productive vocabulary, as indicated by the teacher

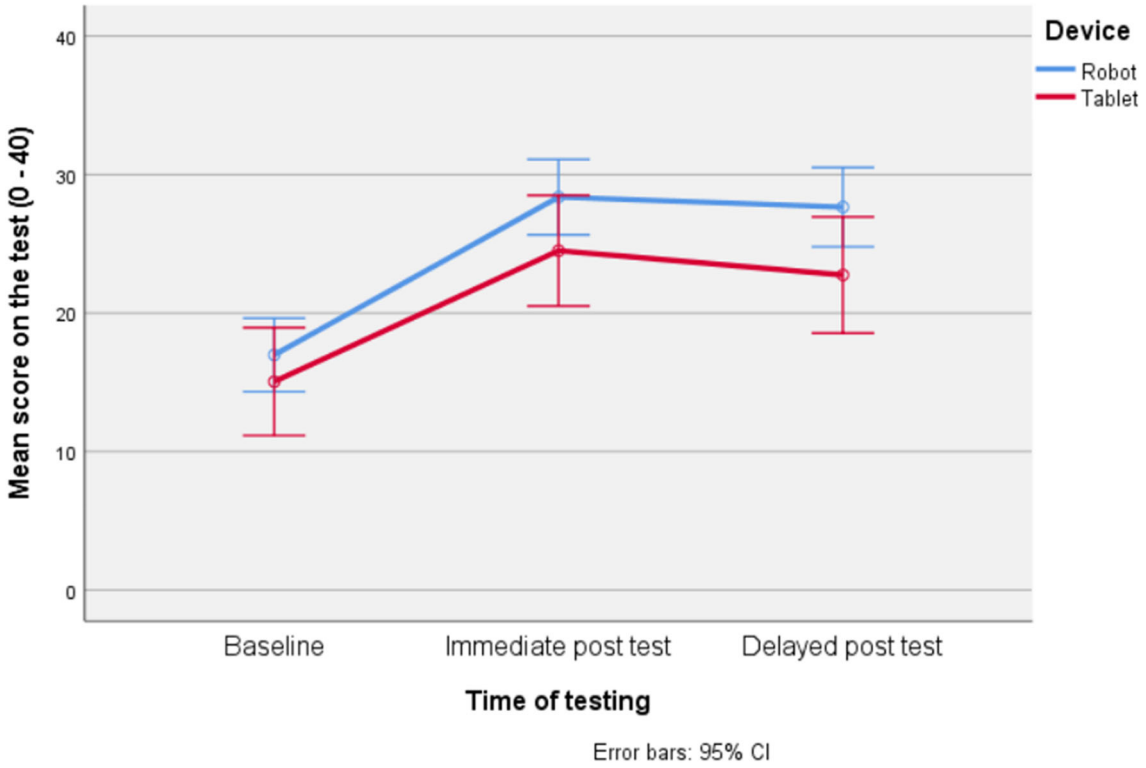


**Fig. 3** Effect of Robot Embodiment on Learning Outcomes. *Note.* Children's mean scores on the vocabulary test with error bars (95% confidence interval) over time (T0: baseline; T1: immediate post-test; T2: delayed post-test), organized by the device they exercised with (blue upper line: robot; red lower line: tablet)

independent variable and perceived humanness scores as dependent variable. Levene's test showed the assumption of equal variances was not violated. Results showed a significant difference in scores of perceived humanness per device ($F(2, 60) = 19.43$, $p < 0.001$, $\eta p^2 = 0.39$), indicating the robot, regardless of behavior, is perceived as more human than the tablet. Social and neutral robots' humanness scores did not significantly differ, thus, behavior of the robot did not affect its perceived humanness. Furthermore, an additional mediation analysis revealed that perceived humanness does not serve as a mediator between device and learning outcomes (see Suppl. Mats., 'Extended analyses H3').

### 3.2.4 Exploratory Analysis

An additional analysis was conducted to determine whether or not behavioral style of the robot affected immediate learning outcomes differently for individual educational differences between children (cf. [54]). To determine the initial educational language advancement, baseline learning outcome scores (T0) were used as basis to divide the sample in two groups depending on whether the individual's baseline performance was above ($N = 22$) or below ($N = 21$) the average of the group. These two groups were included as additional factor in the analysis: a 2 (device) × 2 (advancement) × 2 (time: T0, T1) mixed ANOVA was conducted. For this analysis, it was decided to focus only on the immediate learning outcome effects, both for the sake of simplicity and because the main objective of this analysis was to examine the effect of the robot/tablet on the immediate learning outcomes.

Besides a significant effect for time ($F(1, 39) = 265.57$, $p < 0.001$, $\eta^2 = 0.87$), a significant 3-way interaction for time * device * advancement effect was found ($F(1, 39) = 8.12$, $p = 0.007$, $\eta^2 = 0.17$). Furthermore, a marginally significant device * advancement effect was found ($F(1, 39) = 2.74$, $p = 0.11$, $\eta^2 = 0.07$). After post-hoc testing, children below average language advancement level, exercising with the robot, showed to learn more than children with above average language level, also exercising with the robot (see Fig. 6; also Suppl. Mats., 'Exploratory analyses').

## 4 Discussion

In this study, children were taught new words with a digital tutor, either a robot or a tablet. We tested whether robot tutoring had a more positive effect than tablets on learning outcomes, enjoyment and engagement in (second) language learning with young children, primarily from a migration background. Both devices proved to be adequate tutors as

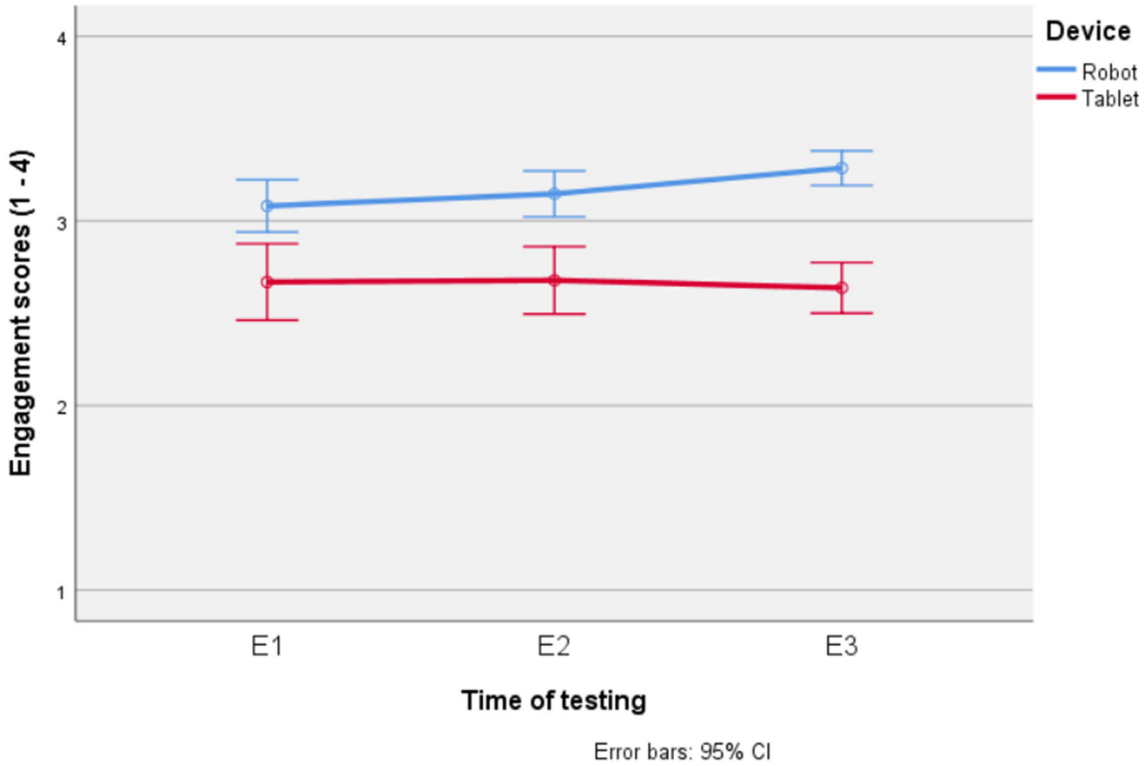


**Fig. 4** Effect of Robot Embodiment on Engagement. *Note.* Children's mean engagement during exercises with error bars (95% confidence interval) over time (E1: first; E2: second; E3: third story-telling exercise), organized by the device they exercised with (blue upper line: robot; red lower line: tablet)

all children improved significantly over time and learned more words. Supporting our assumptions, children who had trained with the robot showed a greater increase in learning outcomes from the baseline test (T0) to the delayed post-test (T2), obtained two weeks after the story-telling exercise. Furthermore, we hypothesized that children would engage more with the robot than the tablet during the story-telling and enjoy the exercises better, which they did. Learning outcomes were also predicted by engagement scores, meaning higher engagement scores stimulated higher learning outcomes. Moreover, children perceived more humanness in the robot than in the tablet. Finally, results showed that the robot's behavioral style did not make a difference overall, that is, whether a robot behaved more socially versus more neutral, just task-oriented, did not affect language learning. However, the behavioral style of the robot did affect children differently according to individual language advancement levels. Children whose word knowledge was below-average starting the experiment, had more to gain from the social behavior of the robot, considerably more than children whose skills were above average and trained with the same socially behaving robot.

To our knowledge, this is the first research in (second) language learning that compares a semi-autonomous embodied robot tutor interacting directly with the learner, without a tablet, to a tablet-only condition. This is an important addition to the body of research thus far, because previous studies combined the robot with a tablet to overcome limitations in ASR. This hindered the potential advantage of robot tutors for seemingly 'face-to-face' interaction. By testing a robot tutor directly interacting with the learner, we gain better insight in the effectiveness of the social robot in comparison to a different technical tutor, the tablet. Our study demonstrated effectiveness of the social robot over time. However, training with the robot, compared to the tablet, did not significantly increase learning outcomes from the baseline (T0) to the immediate post-test (T1), obtained right after the last tutoring session, but was significant at the delayed post-test, 2 weeks later. The relatively large standard deviations make it difficult to reach significance as would be with more homogeneous groups. Our findings are partially in line with previous research, where effectiveness of a social robot in second language learning was shown to be positive [7], although no differences were found in comparison to a tablet [8]. The lack of finding differences thus far was attributed to the robot interacting with the learner via a tablet rather than in direct one-on-one interaction. This is therefore an important contribution of our research, resulting in stronger and lasting effects of language tutoring by a robot over a tablet. Our

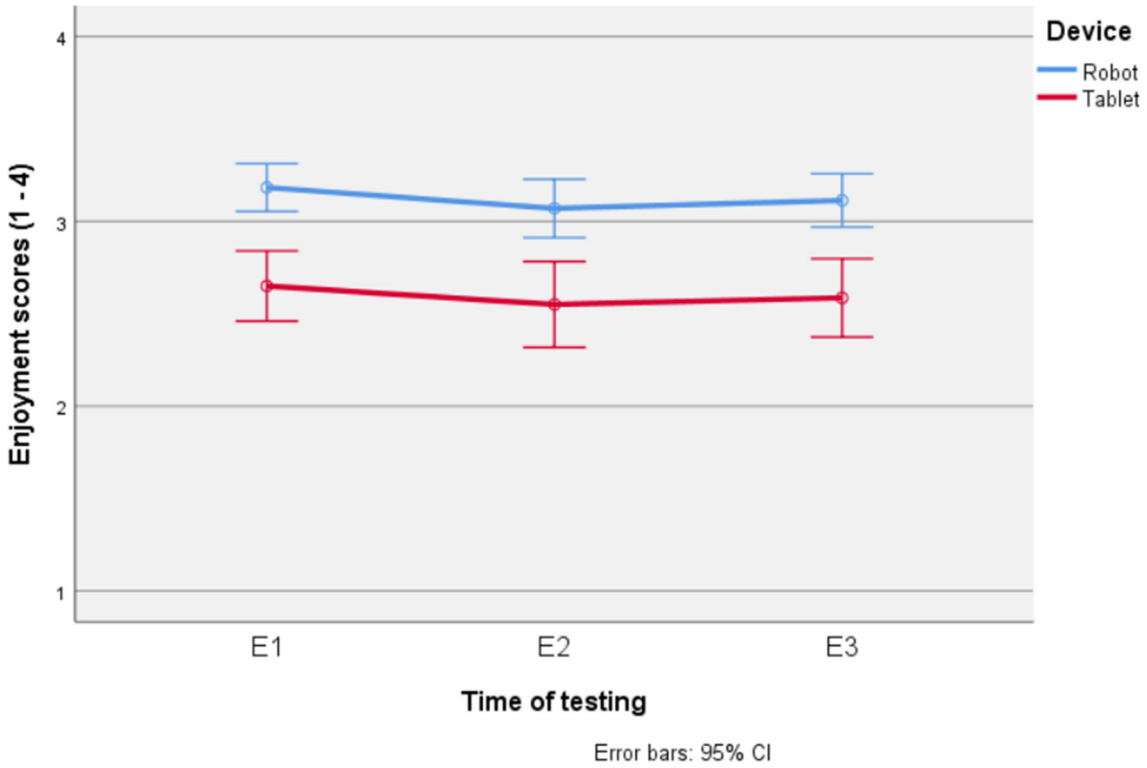


**Fig. 5** Effect of Robot Embodiment on Enjoyment. *Note.* Children's mean enjoyment during exercising with error bars (95% confidence interval) over time (E1: first; E2: second; E3: third story-telling exercise), organized by the device they exercised with (blue upper line: robot; red lower line: tablet)

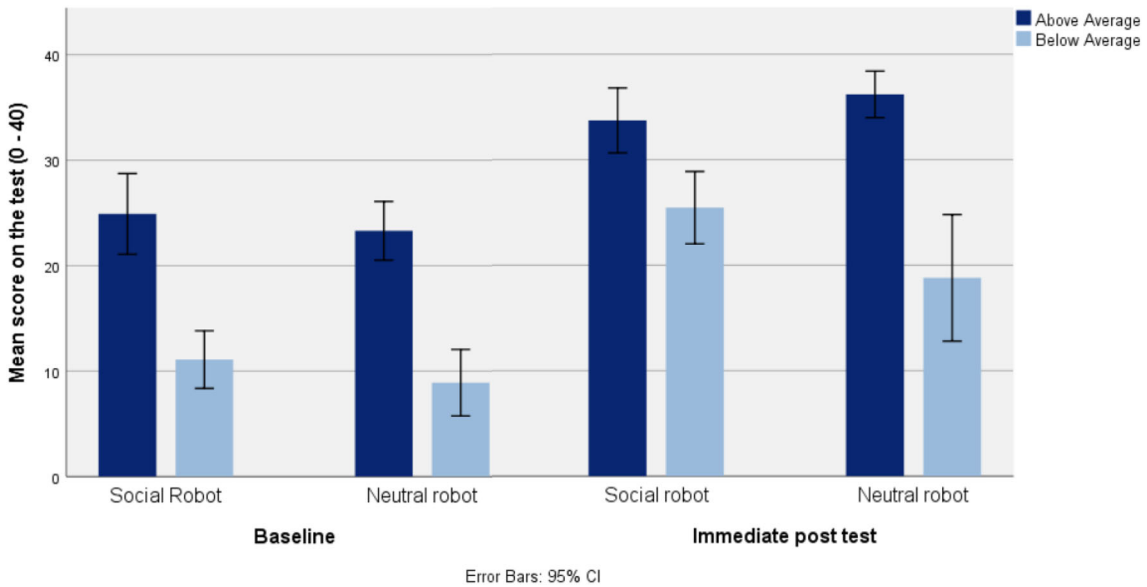


**Fig. 6** The Effect of Robot's behavioral style on learning outcomes based on individual language advancement level. *Note.* Mean learning outcomes on the language test per advancement group, organized by the robot's behavior (i.e., social or neutral), over time (T0: baseline; T1: immediate post-test). Color of bars indicate educational ability level, according to their baseline test scores (dark:'above average'; light:'below average')

results suggest that, over time, children indeed learn more with a robot than a tablet.

Furthermore, the robot significantly increased engagement and enjoyment, and engagement positively affected

learning outcomes. Konishi et al. [35] already stated that interactive environments in which children are engaged, opportunity is created for second language learning and Belpaeme et al. [7] argue that exercising with a robot does just that. In line with previous studies [2, 49, 56], children who trained with a robot engaged more in exercising and enjoyed it better. The distracting nature of the tablet when combined for interaction with the robot (as in [8], might also lower engagement in the task. Moreover, training with a robot might be more captivating, attention-grabbing and lively than training with a tablet, which cannot move nor look at you, and therefore could be perceived as more monotonous and less interesting and enjoyable.

The children in our study perceived the robot as more human than a tablet, which aligns with the notion that robots' humanlike features elicit anthropomorphization [18]. However, in our study, perceived humanness did not predict learning outcomes, in contrast to the findings of Van den Berghe et al. [9], whereas perceived humanness affected engagement and enjoyment positively. This is in line with previous research [53]. While perceived humanness and embodiment (i.e., the use of a robot vs. tablet) both predicted engagement and enjoyment, perceived humanness did not mediate the effect between embodiment and engagement, but did partially mediate the effect between embodiment and enjoyment. Thus, in this study, higher enjoyment in a task can be partially explained by a greater perceived humanness of the device used for the storytelling exercise. This adds to previous research by explaining why enjoyment is higher with a robot [51, 57]. Seemingly, the degree to which children perceived the robot or tablet to be human affected enjoyment, and effectiveness of the robot can be partially explained by its perceived humanness. The high perceived humanness of the robot, alongside its high engagement and enjoyment, thereby creates potential to take up the role of tutor for learning tasks.

The robot was programmed in either a more social or neutral way, but behavioral style of the robot did not affect how many words a child learned overall. Furthermore, whether the robot behaved socially or neutrally did not affect how engaged a child was in working with the robot and how much they enjoyed it. It also did not affect how humanlike they perceived it to be. This lack of different effects of the social versus neutral robot is not a novel finding, mixed results were reported earlier for differently behaving robots in (second) language learning [7]. Using educational ability as a possible explanation to the differential effectiveness of behaviors, Konijn and Hoorn [32] found that in math tutoring the neutral robot was more effective overall, as the social behavior of the robot was perhaps too distracting during such a high-concentration mathematical task. Contrary to those findings, our findings are partially in line with those of Hein and Nathan-Roberts [28], as they support the idea that a socially behaving robot is more successful for language learning, but primarily for less advanced children.

A marginally significant result showed that children who were beyond average in language abilities had more to gain from the neutrally behaving robot than the socially behaving robot. Possibly, above-average children are more task-focused, here learning new words, and the social robot might have distracted them. This study thus showed that individual differences matter in choosing the behavioral style of the robot. The earlier mixed results of the effect of the robot's behavioral style on learning outcomes [7] might be explained through such individual differences and by differences in learning tasks. There seems to be a mechanism that for social tasks, like learning a language, below-average children can benefit from a social style of the robot while above-average children cannot [28]. Whereas for more task-focused learning processes, like math, this mechanism is reversed [32]. Perhaps social behavior matters most for under-achieving students as they require more support in their learning process and in language learning, social behavior might be supportive. It would make sense that social robots "shine" on learning socially relevant skills (such as language), but has less of an effect for purely logical/abstract skills. These findings are certainly very interesting, and future research involving more socially or neutrally behaving robots should take initial knowledge level into account as well as compare different tasks, which might clarify the differential effects of behavior of a social robot on learning outcomes.

Interestingly, a trend emerged for a decrease in learning outcomes from the immediate (T1) to the delayed (T2) post-test. Children who trained with the tablet seemed to have forgotten more words in the time between the last tutoring session and the post-test two weeks later than children who trained with the robot. These results suggest training with a robot possibly enhances retention of knowledge. While this trend was not a significant result, it nonetheless provides an interesting angle of approach for future research.

Furthermore, a negative correlation showed that children who were raised by parents from a (non-western) migration background scored lower on the baseline test than children who only speak Dutch. This is in line with the literature and concerns about bilingual children, and children with parents from a (non-western) migration background lagging behind in Dutch language development [29, 40, 41, 55]. In light of the growing diversity in educational ability levels, cultural backgrounds, and socioeconomic status in classrooms, the number of children who start primary school with a language disadvantage will only increase. Particularly these children stand to benefit from one-on-one tutoring. Embodied robots are good candidates to provide such individual tutoring, something that is currently not achievable for human teachers. Thus, embodied robots carry potential to reduce individual differences and disadvantages.

### 4.1 Limitations and Considerations

For this type of research, the sample size is quite large. However, for the study's power, the small sizes per group in the more complicated statistical analyses are relatively low (i.e., the three-way interactions and split for behavioral style). Moreover, the variance in the test scores appears rather large, which challenges to obtain significance through frequentist statistical testing. Larger sample sizes are hard to obtain in field research involving participants of such a young age, especially when it requires one-to-one interaction in individual sessions. To reach solid conclusions about whether an embodied robot is more suitable for (second) language learning than a tablet, and what possible social behavior works better, additional data collection is needed.

Due to time constraints, children practiced with the robot or tablet 3 times and could not rehearse over a longer time period. Randall [53] recommends that in robot language learning, studies of the effects be at least eight sessions long. Any possible differences will then emerge clearer and make the results more conclusive. Furthermore, in order to implement a fully autonomous social robot into a classroom, the robot should have more technical sophistication. This will also be important to draw a clear distinction between a more social (e.g., speech supportive) and neutral robot.

Furthermore, there were a few outliers (excluded from analyses, see 3.1.1). Because the children were of different ages, some of whom were still really young, their language level may simply not be sufficient to fully comprehend the task and stories. The children with remarkably low scores had a mean age of 4.5 years. The child with the highest score was aged 6.3 years. The task was possibly below level for this child, as this child was in a further stage of language development in comparison to the other participants. This was further confirmed by the language level as scored by their teachers.

In general, robots seem to be effective teaching tools to support human teachers. However, when a robot is actually implemented in educational practice, there will be pedagogical and societal implications involved. Reich-Stiebert and Eyssel [54] captured teacher's concerns and attitudes towards social robots in education. Furthermore, there is still a great deal to unravel before implementing robots in the classroom. Which behavioral style and teaching role (tutor, tutee, peer [16], works best for which task? This research shows that robots can be better tutors than tablets for (second) language learning over time, however, conclusiveness and clarity about the added value of robot's physical embodiment in comparison to other agents needs further study. Importantly, the disadvantages of use of robots in education need not be overlooked, especially for such a young target audience. Tolksdorf et al. [67] mapped ethical concerns of applying robots in Kindergarten settings and thereby emphasize, among other aspects, the vulnerability of children as a group and the role of stakeholders (i.e., teachers and caregivers). Smakman et al. [66] systematically analyzed the moral and ethical considerations of various stakeholder groups that come with the implementation of robot tutors in education. They compared stakeholders' concerns related to the values of friendship and attachment, human contact, privacy, safety, and so on. Concerns need to be extensively researched if robots are to be implemented in (second) language learning.

### 4.2 Conclusions

In all, this research demonstrated effectiveness of a social robot over a tablet over time, but not significantly on immediate learning outcomes. Though both robot and tablet increased word knowledge significantly, robots proved to be more successful than tablets in tutoring over a larger period of time as well as robots being more joyful, engaging and human-like, thereby supporting the learning task. Specifically, our results revealed improved learning outcomes when training with a social robot over time, in contrast to a tablet. Similarly, our results indicate that children who trained with a robot were more engaged in the learning task, and found it more enjoyable. Interestingly, virtually no differences were observed in any of the measures in terms of the robot's behavioral style (social or neutral). However, this (lack of) effect appears to be obfuscated by the educational ability of the child, such that children below average Dutch language ability, were able to learn more from a robot than children with above average Dutch language level interacting with a robot. While social robots need sophistication before being implemented in schools, our study shows the potential of social robots as tutors in (second) language learning.

This study provided a valuable addition to previous research in studying child-robot interaction, without the use of an additional tablet, helping to reveal the true effect of direct one-on-one interaction of social robots in education. While more steps need to be taken before social robots can be implemented in primary education, this study shows that social robot tutors directly addressing the learner has great potential in (second) language learning.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s12369-021-00824-3.

**Acknowledgements** We are very grateful to the schools willing to participate, we particularly thank the teachers and children who voluntarily participated in our study without any further gratification. Furthermore, we would like to thank prof. Dr. Paul Leseman for sharing his insights in developmental literature on language learning, Dr. Paul Vogt for his support in preparing the protocol and the software for the NAO Robot, Max Bode for his support in developing a similar program for the tablet condition, and Zorabots/QBMT for availability of NAO robots.

**Author contributions** Conceptualization: EAK, BJ, VMB; Methodology: EAK, BJ, VMB; Data collection: BJ, VMB; Statistical Analyses: VH, DPV; Visualization: VH, DPV; Funding acquisition: EAK; Project administration: EAK, DPV; Supervision: EAK; Writing – original draft: BJ, VMB, VH, EAK; Writing – review & editing: VH, EAK, DPV, BJ, VMB.

**Funding** This study is funded by the Netherlands Organization for Scientific Research (NWO Open Competition–Digitalization, grant number: 406.DI.19.005).

## Declarations

**Conflict of interest** Authors declare that they have no conflict of interests.

**Data availability** All data, code, and materials used in the analysis are available at OSF<https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/M8SVW1) to researchers for purposes of reproducing or extending the analysis. A brief description of the data set is also provided.



## References

1. Ahmad MI, Mubin O, Shahid S, Orlando J (2019) Robot's adaptive emotional feedback sustains children's social engagement and promotes their vocabulary learning: a long-term child–robot interaction study. Adapt Behav 27:243–266. https://doi.org/10.1177/1059712319844182
2. Ali H, Bhansali S, Köksal I, Möller M, Pekarek-Rosin T, Sharma S, Thebille A-K, Tobergte J, Hübner S, Logacjov A, Özdemir O, Rodriguez Parra J, Sanchez M, Shruti Surendrakumar N, Alpay T, Griffiths S, Heinrich S, Strahl E, Weber C, Wermter S (2019) Virtual or physical? social robots teaching a fictional language through a role-playing game inspired by game of thrones. Soc Robot ICSR 2019(11876):358–367. https://doi.org/10.1007/978-3-030-35888-4_33
3. American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders: DSM-5, 5th edn. DC, Autor, Washington
4. van Baars, L. (2020). De groeiende greep van big tech op het digitale onderwijs [The growing grip of big tech on digital education]. *Trouw*. Retrieved November 12, 2020, from https://www.trouw.nl/onderwijs/de-groeiendegreep-van-big-tech-op-het-digitale-onderwijs~b4cbc7be/
5. Bartneck C, & Forlizzi J (2004) A design-centred framework for social human-robot interaction. In: RO-MAN 2004. 13th IEEE international workshop on robot and human interactive communication (IEEE Catalog No.04TH8759), pp. 591–594
6. 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. https://doi.org/10.1371/journal.pone.0178126
7. Belpaeme T, Kennedy J, Ramachandran A, Scasselati B, Tanaka F (2018) Social robots for education: a review. Sci Robot 3:1–9. https://doi.org/10.1126/scirobotics.aat5954
8. van den Berghe R, Verhagen J, Oudgenoeg-Paz O, van der Ven S, Leseman PPM (2018) Social robots for language learning: a review. Rev Educ Res. https://doi.org/10.3102/0034654318821286
9. van den Berghe R, de Haas M, Oudgenoeg-Paz O, Krahmer E, Verhagen J, Vogt P, Willemsen B, de Wit J, Leseman P (2020) A toy or a friend? Children's anthropomorphic beliefs about robots and how these relate to second-language word learning. J Comput Assist Learn 37:396–410
10. Biemiller A (2012) Teaching vocabulary in the primary grades: vocabulary instruction needed. Guildford Press, New York
11. Bijl, H. (2020). Lerarentekort en coronacrisis, dus staat hier een onbevoegd persoon voor de klas [Teacher deficit and Corona crisis, thus, having an unqualified person in front of the class]. *Het Parool*. Retrieved November 12, 2020, from https://www.parool.nl/amsterdam/lerarentekort-en-coronacrisis-dus-staat-hier-eenonbevoegd-persoon-voor-de-klas~b1eef877/
12. Bloom BS (1984) The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. Educ Res 13(6):4–16
13. Broadbent E (2017) Interactions with robots: the truths we reveal about ourselves. Annu Rev Psychol 68:627–652. https://doi.org/10.1146/annurev-psych010416-043958
14. Cartmill EA, Armstrong BF, Gleitman LR, Goldin-Meadow S, Medina TN, Trueswell JC (2013) Quality of early parent input predicts child vocabulary 3 years later. Proc Natl Acad Sci 110:11278–11283. https://doi.org/10.1073/pnas.1309518110
15. Castellano G, Paiva A, Kappas A, Aylett R, Hastie H, Barendregt W, Nabais F, Bull S (2013) Towards empathic virtual and robotic tutors. Artif Intell Educ 7926:733–736. https://doi.org/10.1007/978-3-642-39112-5_100
16. Chen H, Park HW, Breazeal C (2020) Teaching and learning with children: impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. Comput Educ. https://doi.org/10.1016/j.compedu.2020.103836
17. Davis FD, Bagozzi RP, Warshaw PR (1992) Extrinsic and intrinsic motivation to use computers in the workplace. J Appl Soc Psychol. https://doi.org/10.1111/j.1559-1816.1992.tb00945.x
18. DiSalvo CF, Gemperle F, Forlizzi J, & Kiesler S (2002) All robots are not created equal: the design and perception of humanoid robot heads [ACM]. In: proceedings of the 4th conference on designing interactive systems: processes, practices, methods, and techniques, pp. 321–326
19. Dunn LM, Dunn DM (2007) Peabody picture vocabulary test–4th edition. NCS Pearson, Bloomington
20. Duffy BR (2003) Anthropomorphism and the social robot. Robot Auton Syst 42:177–190
21. Farrel P (2010) School psychology: learning lessons from history and moving forward. Sch Psychol Int 31:581–598. https://doi.org/10.1177/0143034310386533
22. Foster MA, Lambert R, Abbott-Shim M, McCarty F, Franze S (2005) A model of home learning environment and social risk factors in relation to children's emergent literacy and social outcomes. Early Child Res Q 20:13–36. https://doi.org/10.1016/j.ecresq.2005.01.006
23. Golonka EM, Bowles AR, Frank VM, Richardson LD, Freynik S (2014) Technologies for foreign language learning: a review

of technology types and their effectiveness. Comput Assist Lang Learn. https://doi.org/10.1080/09588221.2012.700315
24. Gomez EA, Wu D, Passerini K, Bieber M (2007) Utilizing web tools for computer-mediated communication to enhance team-based learning. Int J Web-Based Learn Teach Technol (IJWLTT) 2:21–37. https://doi.org/10.4018/jwltt.2007040102
25. Gouaillier D, Hugel V, Blazevic P, Kilner C, Monceaux JO, Lafourcade P, & Maisonnier B (2009). Mechatronic design of nao humanoid. In: 2009 IEEE international conference on robotics and automation, pp. 769–774
26. de Haas M, Vogt P, Krahmer E (2020) The effects of feedback on children's engagement and learning outcomes in robot-assisted second language learning. Front Robot AI 7:101. https://doi.org/10.3389/frobt.2020.00101
27. Haßler B, Major L, Hennessy S (2016) Tablet use in schools: a critical review of the evidence for learning outcomes. J Comput Assist Learn 32:139–156. https://doi.org/10.1111/jcal.12123
28. Hein M, Nathan-Roberts D (2018) Socially interactive robots can teach young students language skills; a systematic review. Proc Human Fact Ergon Soc Annu Meet 62:1083–1087. https://doi.org/10.1177/1541931218621249
29. Inspectie van het Onderwijs (2019). De staat van het Onderwijs 2019, retrieved from https://www.onderwijsinspectie.nl/onderwerpen/staat-van-het-onderwijs/documenten/rapporten/2019/04/10/rapport-de-staat-van-het-onderwijs-2019
30. Katsarova, I. (2020). *Teaching careers in the EU: Why boys do not want to be teachers*. https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/642220/EPRS_BRI(2019)642220_EN.pdf
31. 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 (HRI '15), pp. 67–74
32. Konijn EA, Hoorn JF (2020) Robot tutor and pupils' educational ability: teaching the times tables. Comput Educ 157:03970. https://doi.org/10.1016/j.compedu.2020.103970
33. Konijn EA, Hoorn JF (2017) Parasocial interaction and beyond: media personae and affective bonding. In: Rössler P, Hoffner CA, van Zoonen l (eds) The international encyclopedia of media effects. Wiley, London, pp 1–15
34. Konijn EA, Smakman M, van den Berghe R (2020) Use of robots in education. In: van den Bulck J, Sharrer E, Ewoldsen D, Mares M-L (eds) The international encyclopedia of media psychology. Wiley, London, pp 1892–1899
35. Konishi H, Kanero J, Freeman MR, Golinkoff RM, Hirsh-Pasek K (2014) Six principles of language development: implications for second language learners. Dev Neuropsychol 39:404–420. https://doi.org/10.1080/87565641.2014.931961
36. Kory-Westlund JM, & Breazeal CL (2014) Storytelling with robots: learning companions for preschool children's language development. In: 23rd IEEE international symposium on robot and human interactive communication, pp. 643–648
37. Kory-Westlund JM, Dickens L, Jeong S, Harris P, Desteno D, & Breazeal C (2015) 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 extended abstracts, pp. 65–66
38. Laevers F, Daems M, De Bruyckere G, Declercq B, Silkens K, Snoeck G, van Kessel M (2005) Well-being and involvement in care a process-oriented self-evaluation instrument for care settings (SICS). Child & Family and Research Centre for Experiental Education, Leuven
39. van Lehn K (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychologist 46:197–221. https://doi.org/10.1080/00461520.2011.611369
40. Leseman PPM (2000) Bilingual vocabulary development of Turkish preschoolers in The Netherlands. J Multiling Multicult Dev 21:93–112. https://doi.org/10.1080/01434630008666396
41. Leseman PPM, Henrichs LF, Blom E, Verhagen J (2019) Young monolingual and bilingual children's exposure to academic language as related to language development and school achievement. Cambridge University Press, Cambridge
42. Leyzberg D, Spaulding S, Toneva M, & Scassellati B (2014). Personalizing robot tutors to individuals' learning differences. In: Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction (HRI '14), pp. 423–430
43. Mann JA, MacDonald BA, Kuo I-H, Li X, Broadbent E (2015) People respond better to robots than computer tablets delivering healthcare instructions. Comput Hum Behav 43:112–117. https://doi.org/10.1016/j.chb.2014.10.029
44. Marulis LM, Neuman SB (2010) The effects of vocabulary intervention on young children's word learning: a meta-analysis. Rev Educ Res 80:300–335. https://doi.org/10.3102/0034654310377087
45. Mol SE, Bus AG, Jong MT, d., & Smeets, D. J. H. (2008) Added value of dialogic parent–child book readings: a meta-analysis. Early Educ Dev 19:7–26. https://doi.org/10.1080/10409280701838603
46. Moore JB, Yin Z, Hanes J, Duda J, Gutin B, Barbeau P (2009) Measuring enjoyment of physical activity in children: validation of the physical activity enjoyment scale. J Appl Sport Psychol 21:s116–s129. https://doi.org/10.1080/10413200802593612
47. Moriguchi Y, Okanda M, Itakura S (2008) Young children's yes bias: How does it relate to verbal ability, inhibitory control, and theory of mind? First Lang 28:431–442. https://doi.org/10.1177/0142723708092413
48. Otterborn A, Schönborn K, Hultén M (2019) Surveying preschool teachers' use of digital tablets: general and technology education related findings. Int J Technol Des Educ. https://doi.org/10.1007/s10798-018-9469-9
49. Park HW, Grover I, Spaulding S, Gomez L, & Breazeal C (2019) A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In: The thirty-third AAAI conference on artificial intelligence (AAAI-19), vol. 33
50. Paslwaski T (2005) The clinical evaluation of language fundamentals, fourth edition (CELF-4): a review. Can J Sch Psychol 20:129–134. https://doi.org/10.1177/0829573506295465
51. Pereira, A., Martinho, C., Leite, I., & Paiva, A. (2008). I-Cat, the chess player: the influence of embodiment in the enjoyment of a game. In: Proceedings of the 7th international joint conference on autonomous agents and multiagent systems, vol. 3 pp. 1253–1256
52. Pulido JC, González JC, Suárez-Mejías C, Bandera A, Bustos P, Fernández F (2017) Evaluating the child–robot interaction of the NAO therapist platform in pediatric rehabilitation. Int J Soc Robot 9:343–358. https://doi.org/10.1007/s12369-017-0402-2
53. Randall N (2019) A survey of robot-assisted language learning (rall). ACM Trans Human-Robot Interact 9:36. https://doi.org/10.1145/3345506
54. Reich-Stiebert N, Eyssel F (2016) Robots in the classroom: what teachers think about teaching and learning with education robots. In: Agah A, Cabibihan JJ, Howard A, Salichs M, He H (eds) Social Robotics. ICSR 2016. Lecture notes in computer science, vol 9979. Springer, Cham
55. Scheele AF, Leseman PPM, Mayo AY (2010) The home language environment of monolingual and bilingual children and their language proficiency. Appl Psycholinguist 31:117–140. https://doi.org/10.1017/S0142716409990191
56. Schodde T, Hoffmann L, Stange S, Kopp S (2019) Adapt, explain, engage—a study on how social robots can scaffold second-

language learning of children. ACM Trans Human Robot Interact. https://doi.org/10.1145/3366422
57. Sinoo C, van der Pal S, Blanson Henkemans OA, Keizer A, Bierman BPB, Looije R, Neerincx MA (2018) Friendship with a robot: children's perception of similarity between a robot's physical and virtual embodiment that supports diabetes self-management. Patient Educ Couns 101:1248–1255. https://doi.org/10.1016/j.pec.2018.02.008
58. Tamim, R. M., Borokhovski, E., Pickup, D., Bernard, R. M., & El Saadi, L. (2015). Tablets for teaching and learning: a systematic review and meta-analysis [report] [Retrieved from Commonwealth of Learning (COL), website: http://oasis.col.org/handle/11599/1012]
59. Valli A (2008) The design of natural interaction. Multimed Tools Appl 38:295–305. https://doi.org/10.1007/s11042-007-0190-z
60. Vogt, P., van den Berghe, R., de Haas, M., Hoffmann, L., Kanero, J., Mamus, E., & Pandey, A. K. (2019). Second language tutoring using social robots. A large-scale study. In: IEEE/ACM International Conference on Human-Robot Interaction (HRI 2019). https://pub.uni-bielefeld.de/record/2933181
61. Vygotsky LS (1980) Mind in society: the development of higher psychological processes. Harvard University Press, London
62. Wayne AJ, Youngs P (2003) Teacher characteristics and student achievement gains: a review. Rev Educ Res 73:89–122. https://doi.org/10.3102/00346543073001089
63. de Wit, J., Brandse, A., Krahmer, E., & Vogt, P. (2020). Varied human-like gestures for social robots: Investigating the effects on children's engagement and language learning, In: Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction, Cambridge, United Kingdom, Association for Computing Machinery, doi: https://doi.org/10.1145/3319502.3374815
64. Schmider E, Ziegler M, Danay E, Beyer L, Bühner M (2010) Is it really robust? Methodology 6:147–151. https://doi.org/10.1027/1614-2241/a000016
65. Skinner EA, Belmont MJ (1993) Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. J Educ Psychol 85:571. https://doi.org/10.1037/0022-0663.85.4.571
66. Smakman MHJ, Konijn EA, Vogt P, Pankowska P (2021) Attitudes towards social robots in education: enthusiast, practical, troubled, sceptic, and mindfully positive. Robotics 10:24. https://doi.org/10.3390/robotics10010024
67. Tolksdorf NF, Siebert S, Zorn I, Horwath I, Rohlfing KJ (2020) Ethical considerations of applying robots in kindergarten settings: towards an approach from a macroperspective. Int J Soc Robot. https://doi.org/10.1007/s12369-020-00622-3