International Journal of Child-Computer Interaction 33 (2022) 100501

Contents lists available at ScienceDirect

International Journal of Child-Computer Interaction

journal homepage: www.elsevier.com/locate/ijcci

Research paper

Engagement in longitudinal child-robot language learning interactions: Disentangling robot and task engagement☆

Mirjam de Haas [a,b,*], Paul Vogt [a,c], Rianne van den Berghe [d], Paul Leseman [e], Ora Oudgenoeg-Paz [e], Bram Willemsen [a,f], Jan de Wit [g,b], Emiel Krahmer [g,b]

[a] Department of Cognitive Science and Artificial Intelligence, Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg, the Netherlands
[b] Tilburg center for Cognition and Communication, Tilburg University, Tilburg, the Netherlands
[c] School of Communication, Media & IT, Hanze University of Applied Sciences, Groningen, the Netherlands
[d] Section Leadership in Education and Development, Windesheim University of Applied Sciences, Almere, the Netherlands
[e] Department of Development of Youth and Education in Diverse Societies, Utrecht University, Utrecht, the Netherlands
[f] Department of Intelligent Systems, KTH Royal Institute of Technology, Stockholm, Sweden
[g] Department of Communication and Cognition, Tilburg University, Tilburg, the Netherlands

ARTICLE INFO

Article history:
Received 9 July 2021
Received in revised form 3 April 2022
Accepted 23 May 2022
Available online 9 June 2022

Keywords:
Child-robot interaction
Engagement
Second language learning
Robot tutor
Preschool children

ABSTRACT

This study investigated a seven sessions interaction between a peer-tutor robot and Dutch preschoolers (5 years old) during which the children learned English. We examined whether children's engagement differed when interacting with a tablet and a robot using iconic gestures, with a tablet and a robot using no iconic gestures and with only a tablet. Two engagement types were annotated (task engagement and robot engagement) using a novel coding scheme based on an existing coding scheme used in kindergartens. The findings revealed that children's task engagement dropped over time in all three conditions, consistent with the novelty effect. However, there were no differences between the different conditions for task engagement. Interestingly, robot engagement showed a difference between conditions. Children were more robot engaged when interacting with a robot using iconic gestures than without iconic gestures. Finally, when comparing children's word knowledge with their engagement, we found that both task engagement and robot engagement were positively correlated with children's word retention.

1. Introduction

Engagement is important for learning (Christenson, Wylie, & Reschly, 2012; Zaga, Lohse, Truong, & Evers, 2015). The more time children are actively interacting with a certain task the more children can learn. It can also increase children's motivation. When children stay engaged, they are motivated to learn, actively use their newly gained knowledge and will continue the learning session or even try more challenging tasks, which can lead to higher learning gains (Jang, 2008).

The large role of engagement in learning is one of the reasons why engagement is a well known concept in the field of educational human–robot interaction (HRI) (van den Berghe, Verhagen, Oudgenoeg-Paz, van der Ven, & Leseman, 2019; Kanero et al., 2018). Children are generally highly engaged with robots, however most studies only include short-term interventions with a robot (van den Berghe et al., 2019). Therefore, high engagement of children might also be a result of a novelty effect (i.e. the -often exciting- effect that interacting with a novel technology can have on engagement) (Kanda, Hirano, Eaton, & Ishiguro, 2004; Leite, Martinho et al., 2013). The few studies that investigated children's engagement during a longer period noticed that children's engagement started to decline fairly quickly after a few sessions (Ahmad, Mubin, & Orlando, 2017; Kanda, Sato, Saiwaki, & Ishiguro, 2007; Komatsubara, Shiomi, Kanda, Ishiguro, & Hagita, 2014; Leite, Martinho et al., 2013).

It is important to bear in mind that most long-term HRI studies that studied engagement, only investigated engagement

☆ **Acknowledgment**
This work has been supported by the EU H2020 L2TOR project (grant 688014). We would like to thank the children, their parents, and the schools for their participation. Furthermore, we would like to thank Laurette Gerts, Annabella Hermans, Esmee Kramer, Madee Kruijt, Marije Merckens, David Mogendorff, Sam Muntjewerf, Reinjet Oostdijk, Laura Pijpers, Chani Savelberg, Robin Sonders, Sirkka van Straalen, Sabine Verdult, Esmee Verheem, Pieter Wolfert, Hugo Zijlstra and Michelle Zomers for their help in collecting data and annotating videos.

* Corresponding author at: Department of Cognitive Science and Artificial Intelligence, Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg, the Netherlands.
E-mail address: mirjam.dehaas@tilburguniversity.edu (M. de Haas).

https://doi.org/10.1016/j.ijcci.2022.100501

in general. Often this means that these studies investigated the engagement between robot and user, as interactions are a social process. However, children can also engage with a task in front of them, instead with only their social partner. Therefore, it has become increasingly more apparent that there should be a distinction between the engagement with the task (*task engagement*) and engagement between the learner and the robot (*robot engagement*) (Zaga, Truong, Lohse, & Evers, 2014).[1]

It is still unclear whether task engagement and robot engagement have a positive or a negative relation with learning gain. Although one might expect that a robot behaving in a way that stimulates engagement (a higher robot engagement) leads to better learning outcomes, it is also possible that a more engaging robot will distract the child, which can result in the child paying less attention to the learning task in front of them (Kennedy, Baxter, & Belpaeme, 2015). Instead, children interacting with a less distracting robot might pay more attention to the task, become more task-engaged and less robot-engaged but learn more because they have more attention for the task.

Therefore, it can be important to look closely at the difference between children's engagement with a robot and the task and whether a decline in engagement is something specifically related to the robot, the robot's behavior or a more general effect of sessions with technological devices. After all, children's (task) engagement might also drop when they are interacting with only a tablet. In addition to the physical presence of the robot, it is possible that the non-verbal behavior of a robot, such as the use of gestures or head movements, might lead to a higher level of engagement than other electronic devices, including tablets or computers. This non-verbal behavior of a robot, such as the use of gestures, can benefit children's learning gain. This was found especially in second-language (L2) learning in which gestures can be used not only to direct students' attention, but also as scaffolding techniques (de Wit et al., 2018). Scaffolding can be provided by using gestures that depict the meaning of a concept (iconic gestures), which can strengthen the connection between the new L2 concept and the known first-language word (Roth, 2001). These types of gestures are shown in a previous study to benefit children's engagement during a short-term interaction (de Wit et al., 2018). Whether this effect remains during multiple sessions is something we will explore in the current study.

This study is carried out within the L2TOR project[2], a project that investigated how a robot can teach pre-school children a second language. In this article, we studied the effects of engagement in an L2 learning setting. An advantage of studying engagement in an L2 setting is that this provides a perfect situation for having social interactions between child and robot in a clear task environment. Moreover, it offers the opportunity to investigate the effect of a robot's gesture use and the relation of children's engagement and their L2 word knowledge.

2. Background

2.1. Robots in education

Social robots have been used in education for quite some time now (for a review, see Belpaeme, Kennedy et al., 2018). They have been used with children in many fields, such as teaching children mathematics or helping children with writing (Alves-Oliveira, Sequeira, Melo, Castellano, & Paiva, 2019; Kennedy et al., 2015; Konijn & Hoorn, 2020) but also supporting children when learning a second language (e.g., Belpaeme, Vogt et al., 2018; Konijn, Jansen, Mondaca Bustos, Hobbelink, & Preciado vanegas, 2021; Kory-Westlund & Breazeal, 2015), which is the context of this article.

Most interactions between robots and children rely on other devices, such as tablets, to get an autonomous interaction since speech recognition has shown to still be unreliable (Kennedy et al., 2017; Mubin et al., 2012) and that technical breakdowns negatively impact the interaction (Salichs et al., 2019). To deal with this shortcoming, some child-robot interactions rely on the improvement of speech recognition in the coming years and use a Wizard of Oz approach for their studies (e.g., Kory-Westlund et al., 2017). Other researchers use an extra device, such as a tablet as input instead of relying on speech commands. The added advantage of using a tablet is that the display can create a virtual environment for the interaction and the robot can manipulate things on the tablet more easily than in the physical world. For example, a tablet screen has been used to give children the impression that a robot is able to write (Jacq, Lemaignan, Garcia, Dillenbourg, & Paiva, 2016), to display an interactive city map and use the screen for turn-taking between children and the robot (Alves-Oliveira et al., 2019) or to display a board game (Snake and Ladders) that children played with the robot in order to improve the robot's autonomous behavior. (Ahmad, Mubin, Shahid, & Orlando, 2019). In all these studies, the tablet was used in order to facilitate an autonomous interaction between child and robot, by either giving the impression that the robot can write on a screen or by using the input by the child on the screen to provide the robot with the current game state. However, these studies did not investigate the added advantage of the robot compared to only a tablet. It is interesting to examine whether the presence of the robot is not distracting the child from the task, and whether engagement with the robot assists in learning.

Moreover, the advantage of the robot's presence rather than only a tablet or computer is to enable children to interact more naturally with a robot than with a computer screen or tablet since a robot makes use of non-verbal behavior, such as using its arms for gesturing or nodding its head for confirmation (van den Berghe et al., 2019; Kory-Westlund et al., 2015). These gestures can be used for scaffolding, and can support grounding of the unknown L2 concept in the familiar language. The use of iconic gestures, gestures that depict the meaning of a certain concept, can support L2 learning in human–human studies (Macedonia, Müller, & Friederici, 2011), and in short-term child-robot interactions (de Wit et al., 2018). However, these studies did not compare a robot with a tablet, nor examined long-term effects. Our study hopes to provide further insights into the effect of a robot's presence and the robot's use of gestures, using the tablet as a learning device, on the child's engagement.

2.2. Engagement

Despite its common usage, there are multiple definitions of engagement used for HRI. The definition by Sidner and colleagues (Sidner, Lee, Kidd, Lesh, & Rich, 2005) is the most commonly-used definition in HRI (Oertel et al., 2020). Sidner and colleagues defined engagement as "the process by which individuals in an interaction start, maintain and end their perceived connection to one another" (p. 141). This definition mostly focuses on the cognitive aspect of the individuals who are interacting. When we investigate engagement with learning, or school engagement, the construct engagement becomes more than merely a cognitive engagement between interaction partners. Rather, it relates to an interaction between an emotional dimension, a cognitive dimension, and a behavioral dimension (Fredricks, Blumenfeld,

[1] The term robot engagement is commonly referred to as social engagement. We prefer to use the term robot engagement to clearly indicate that we refer to engagement between robot and child because social engagement can also include interactions between the child and other actors, such as the experimenter in the room, or other children in the case of a group interaction.

[2] see www.l2tor.eu

& Paris, 2004). The *emotional* dimension relates to how children feel during the session and to what extent they like the robot; the *cognitive* dimension relates to the effort the child puts into the task and how determined the child is to succeed. Finally, the *behavioral* aspect involves the attentiveness of the child, for example to what extent the child pays attention to the task and how the child responds to the instructions. When referring to engagement, it is this complex, multidimensional construct that incorporates all of these dimensions.

As a consequence of the complexity, it is challenging to measure engagement and previous published work in HRI is not consistent in the way of measuring engagement. Many studies focus on a single aspect of engagement, such as eye gaze and speech which are elements of the cognitive aspect of engagement (Chaspari, Al Moubayed, & Fain Lehman, 2015; Chung, 2019; Xu, Zhang, & Yu, 2016), often in combination with behavioral aspects such as smiles and nods (Serholt & Barendregt, 2016), gestures (Ahmad et al., 2019; Tapus et al., 2012) and initiations by the child (Javed, Jeon, & Park, 2018; Tapus et al., 2020). There are a few disadvantages of using only these measurements. Using eye gaze, for example, overlooks the fact that when children do not look at the robot, this does not necessarily imply that children are not engaged with the interaction. Sometimes children need to look at the task in front of them instead of the robot, while being engaged. Likewise, for children's speech, when the child-robot task requires them to use speech, it is still possible that children are answering the question in order to continue with the task without being actively task-engaged. Other studies measure the body posture of the child and the distance to the robot (Heath et al., 2017; Sanghvi et al., 2011), sometimes in combination with speech (Javed, Lee, & Park, 2020; Jeong, Breazeal, Logan, & Weinstock, 2018) or in combination with touch behavior on a tablet (Vázquez, Steinfeld, Hudson, & Forlizzi, 2014). However, these studies do not take into account that the task might require children to move around and children are in general more active and move around more than adults.

Moreover, other studies use measurements that are not suitable for younger children, such as a questionnaire (Díaz, Nuño, Saez-Pons, Pardo, & Angulo, 2011; Zaga et al., 2015), sometimes combined with other techniques such as the use of distractors (Ligthart, Neerincx, & Hindriks, 2020) or physiological measurements such as thermal infrared imaging (Filippini et al., 2020), or electrodermal activity (Leite, Henriques et al., 2013) and EEG (Alimardani & Hiraki, 2020; Perugia, Díaz-Boladeras, Català-Mallofré, Barakova, & Rauterberg, 2020; Szafir & Mutlu, 2012). These measurements are not only invasive for children, but can also introduce more overhead during the experiment which makes it more difficult to use outside the lab.

There have been few studies in which engagement was detected automatically (Ishii & Nakano, 2010; Rich, Ponsleur, Holroyd, & Sidner, 2010; Rudovic, Lee, Dai, Schuller, & Picard, 2018), however it is difficult to be certain these automatic measurements are actually measuring engagement. Often they are based on only one dimension like verbal utterances, or emotional features. Or they are based on deep learning, which needs a lot of data to be reliable (Rudovic et al., 2018) which makes it less feasible to use for every study. Automatic engagement measurements are additionally sensitive for errors because of the focus on one dimension and they are also more susceptible to error because they are automated.

The main limitation of all these different measurements is that they do not provide a complete overview of children's engagement but rather a one-sided aspect of engagement. Additionally, these studies did not take into account that there are differences between the engagement between child and robot, and engagement between child and task. A child can be very engaged with the task and only focusing on the task, while not being engaged with the robot or vice versa. It is therefore important to make a distinction between engagement with the task, and engagement with the robot (Oertel et al., 2020; Zaga et al., 2015).

Zaga et al. (2015) specifically investigated task engagement. They compared the puzzle solving ability of children between 6 and 9 years old with a robot that behaved either in a peer-like or a tutor-like manner. They measured children's gaze as part of the cognitive component of engagement, the children's puzzle completion for their behavioral component and they used a questionnaire to measure children's emotion for the task. They found that children were more task-engaged with the peer-like robot than with a tutor-like robot, and could solve the puzzles faster when interacting with the peer-like robot. However, this interaction was only one session and this makes it difficult to generalize the results to multiple sessions.

With respect to the level of engagement, it may seem that higher engagement is always preferable but robot engagement can also have negative outcomes on children's learning performance. For example, Kennedy and colleagues (Kennedy et al., 2015) reported that children more focused on the robot (which might indicate high robot engagement) scored lower than children less focused on the robot. This particular study investigated children's mathematical skills and did not focus on children's language skills so whether this can be generalized to L2 learning has yet to be confirmed. It is possible that language learning depends more on interaction between partner and participant and that a high robot engagement will have a positive influence on children's L2 learning outcome. Kennedy and colleagues' outcomes provide the indication that robot engagement is not always the single element for an interaction to be successful involving learning.

We propose to measure both task engagement and robot engagement based on children's video observations, with the use of a grounded coding scheme called ZIKO (Laevers, 2005), which combines different aspects of engagement.

2.2.1. ZIKO[3]

The ZIKO observation instrument is a method that has been used to observe children in kindergarten during their daily activities. The scheme is based on developmental schemes (Laevers, 2005) to create a 5-point Likert scale that rates multiple aspects of children's behavior, such as the well being of the child, but also engagement of the child. The scheme has been used to improve activities at kindergartens (e.g., Arnott, Grogan, & Duncan, 2016; Laevers, 2015; Storli & Sandseter, 2019) and to get an evaluation of a particular child or the activities played by the children (Storli & Sandseter, 2019) and has been shown to be relatively stable (Laevers, 2015). The scheme has additionally been used in research more related to child-robot interaction: to compare children's engagement with an iPad versus children's creative play (Arnott et al., 2016).

The engagement component of the instrument is a detailed scheme that includes the three components of engagement proposed (Fredricks et al., 2004): children's levels of concentration, motivation (cognitive dimension), energy (emotional dimension), their exploratory drive and persistence (behavioral dimension) and when all these components are present in children's behavior, children are highly engaged. The main advantages of using this scheme is that it provides a score, which allows for quantitative analyses over time and it has been designed for preschool children.

[3] ZIKO is an abbreviation for Zelfevaluatie-Instrument voor de Kinderopvang (English: Self-evaluation Instrument for Care Settings).

2.3. Long-term interactions

Long-term interactions are important to investigate because they look beyond the novelty effect (Ahmad et al., 2019; Kanda et al., 2007; Leite, Martinho et al., 2013; Oertel et al., 2020). Salter, Dautenhahn, and Bockhorst (2004) suggested that you can speak of long-term interaction after the novelty effect is gone and the experimenters are left with an interaction between robot and child without any interference of the novelty effect. In their study, children did not show any interest in the robot anymore after three sessions in the case that the robot used repeated behavior. This can also be confirmed by Serholt and Barendregt (2016) who found that children's social responses to the robot were drastically reduced by the third session. Three other studies investigated primary school children (8–9 years) over time (Ahmad et al., 2019; Davison et al., 2020; Leite, Martinho et al., 2013) and found that the children's engagement remained the same over time when playing chess five times during five weeks (Leite, Martinho et al., 2013) or over three sessions when the robot was adapting itself to the child's emotional state during a second-language learning task (Ahmad et al., 2019). Davison et al. (2020) used an autonomous robot to practice science-related tasks with children (6–10 years old) for four months. They found that children's interest in the lessons dropped but that it increased again when the robot started to discuss new materials. These studies focused on older children at primary schools and children at that age undergo major developmental changes, which results that there are large learning differences between older children and younger children (Piaget, 1976). Considering the fact that children are more likely to learn a language at a young age, it would therefore be worthwhile to include younger children.

Three long-term studies that investigated younger children were de Haas, Krahmer, and Vogt (2020), Kanda et al. (2007), Tanaka, Cicourel, and Movellan (2007). Kanda et al. (2007) placed a robot in a preschool during two months and found that children's initial social bond with the robot seems to relate with their robot engagement. Children who established a social bond with the robot, continued the interaction for a longer period than children who did not have this social bond. Moreover, Tanaka et al. (2007) showed that children's engagement quickly decreased and that only after introducing new robot behaviors, children returned to the robot. These two interactions were free play interactions, meaning that the robot was more a playmate than a tutor and the question remains whether children's engagement and learning gain are related. In de Haas et al. (2020), it was found that a robot providing teacher-like feedback had a positive influence on children's task engagement and robot engagement. However, this study only contained three sessions and the question remains what will happen to children's task engagement and robot engagement after more sessions when the novelty plays a smaller role.

2.4. This study

The current study was part of a large-scale study in which we investigated the effectiveness of a peer-tutor robot (Softbank Robotics NAO robot) in a long-term L2 tutoring interaction, teaching pre-school children some English vocabulary as second language (Vogt et al., 2019). This study's experimental design, hypotheses and statistical analyses were preregistered on AsPredicted [4] and source code has made publicly available via Github [5]. The study included four conditions: (1) an L2 tutoring training with a tablet and a robot using iconic gestures (gestures that act out the meaning of a word) and deictic gestures (pointing gestures), (2) an L2 tutoring training with a robot using deictic gestures, (3) an L2 tutoring training with a tablet, and (4) a control condition in which children danced with the robot but were not taught any English words. Word knowledge was tested on three occasions: a pre-test, an immediate post-test and a delayed post-test (administered between two and four weeks after the last session). The results of the preregistered study were presented in Vogt et al. (2019) and showed that children scored higher after the tutoring sessions than before. Moreover, children in the experimental conditions (robot with iconic gestures, robot without iconic gestures, tablet-only condition) scored significantly higher than children in the control condition on the immediate and delayed post-test. There were no significant differences between the experimental conditions in children's English word knowledge, meaning that children in the robot conditions did not learn more than in the tablet-only condition.

In this current paper, we present the first longitudinal comparison of robot engagement and task engagement, the role of iconic gestures and the presence of a physical robot, and its link to second-language word knowledge. We measured children's task engagement, their robot engagement and children's L2 word knowledge to investigate the relation between engagement and L2 word knowledge. In this paper, we only included the three experimental conditions because the control condition interaction in the original study was very different from the other three conditions. We addressed the following hypotheses:

H1 Task engagement

(a) Children are more *task-engaged* when interacting with a robot and a tablet compared to a tablet only
(b) Children's *task engagement* decreases less over time when interacting with a robot and a tablet compared to a tablet only

H2 Robot engagement

(a) Children are more *robot-engaged* with a robot using iconic gestures than one without iconic gestures
(b) Children's *robot engagement* decreases less with a robot using iconic gestures than one without iconic gestures

H3 Relation engagement and word-knowledge

(a) Children's task engagement is positively related with children's L2 word knowledge
(b) Children's robot engagement is negatively related with children's L2 word knowledge, based on the results by Kennedy et al. (2015)

3. Method

3.1. Participants

We recruited 208 native Dutch speaking children from nine different Dutch primary schools. The children's mean age was 5 years and 8 months (SD = 5 months). All parents gave informed consent. Three children were excluded due to a high prior-knowledge of English as measured in the pre-test. During the experiment nine children dropped out due to various reasons, such as sickness or experiment anxiety and two children were excluded due to technical errors. This resulted in a total of 194 children. The children were pseudo-randomly (taking their pre-test score and gender into account) assigned to one of the four conditions:

1. Robot with iconic gestures: $N = 54$, $M_{age} = 5$ years and 8 months, $SD = 5$ months, 31 boys and 23 girls

[4] see https://aspredicted.org/6k93k.pdf
[5] see https://github.com/l2tor

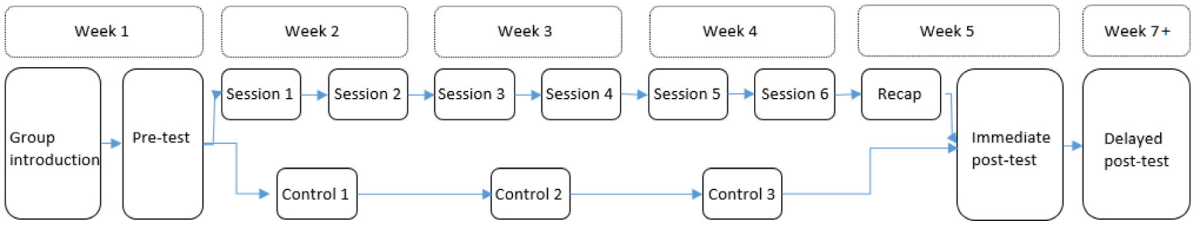


Fig. 1. Schematic overview of the experiment.

2. Robot without iconic gestures: $N = 54$, $M_{age} = 5$ years and 8 months, $SD = 5$ months, 28 boys and 26 girls
3. Tablet-only: $N = 54$, $M_{age} = 5$ years and 9 months, $SD = 5$ months, 24 boys and 30 girls
4. Control: $N = 32$, $M_{age} = 5$ years and 7 months, $SD = 5$ months, 14 boys and 18 girls

The study was conducted in accordance with the Declaration of Helsinki. The project in which the study was embedded, the L2TOR project, received ethical approval from Utrecht University's Ethics Committee under protocol number FETC16-039.

3.2. Design

The experiment consisted of a pre-test, seven tutoring sessions (with the final one being a recap session), an immediate post-test and a delayed post-test (a schematic overview can be found in Fig. 1). It was a between-subjects design where children received tutoring sessions with a robot (using iconic gestures or no iconic gestures) and a tablet or only with a tablet. These tutoring sessions were completely the same except for the physical presence of the robot and the use of iconic gestures. In the robot with iconic gestures condition, the robot used an iconic gesture every time it said an L2 target word and it used deictic gestures such as pointing when children had to perform a task on the tablet. In the robot without iconic gestures, the robot only used deictic gestures, and no iconic gestures. In the tablet-only condition, children heard the voice of the robot through the tablet's speakers, but did not see the robot's physical presence during the experiment. Children in the control condition received three one-on-one sessions with a robot without any English tutoring, participating in dancing activities instead.

We only measured task engagement for the experimental conditions (robot with iconic gestures, robot without iconic gestures and tablet-only) because these groups were participating in the tutoring sessions. Robot engagement was only measured for the experimental conditions with a robot present (robot with iconic gestures and robot without iconic gestures). The control condition was not included in this study, because this interaction was very different than the other interactions and the specific interest of this study is children's engagement and the relation with learning and the children in the control condition did not receive the learning activities.

3.3. L2 tutoring sessions

The aim of the L2 tutoring sessions was to teach each child 34 English words. Each child received seven sessions with the robot and a tablet or only the tablet (see for an example Fig. 2). Children were taught approximately six target words during each session, except the seventh session which was a recap session. Children heard the new target words ten times during the session, and these new target words were repeated once during the following session and twice during the recap session. The target words can be found in Table 1, which were divided in two domains: number domain and spatial domain. The number domain consisted of

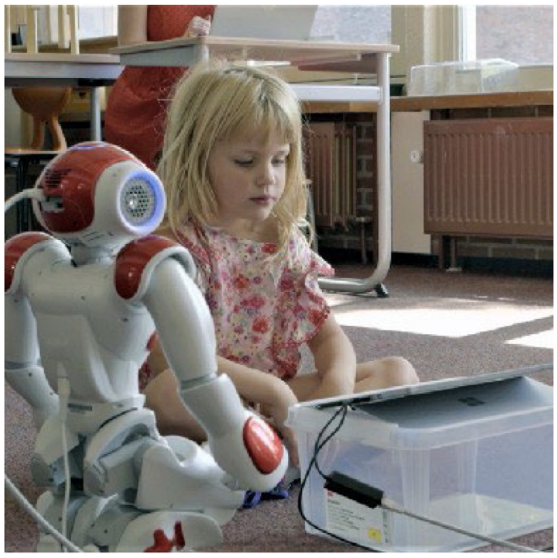


Fig. 2. Experimental setting.

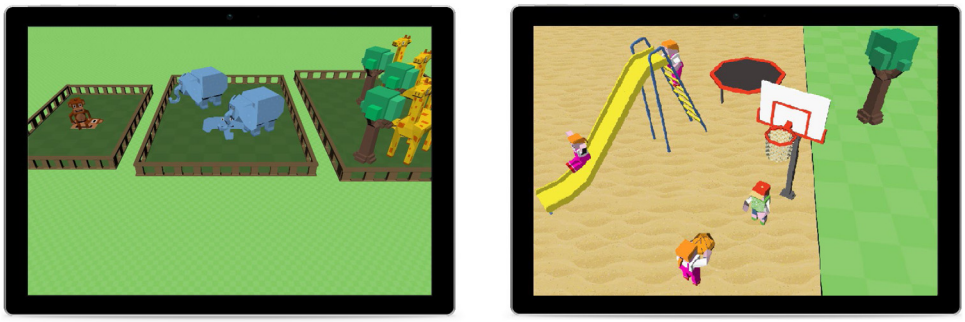
counting words (e.g., one, two), verbs for mathematical operations (e.g., adding and take away) and comparisons (e.g., more, most). The spatial domain contained prepositions (e.g., in front of, on) and action verbs (e.g. running, climbing). Each session was presented in a different virtual environment that was designed to teach the target words specific for that session, for example in session one (see Fig. 3a), each of the cages contained different amounts of animals and after the animals escaped their cages, children had to return (*add*) the animals to their cages. Similarly, in session six (see Fig. 3b), the tablet displayed a child sliding and climbing a slide.

Each of the tutoring sessions followed a similar script, and contained a few personalized interactions such as the use of the child's name in the beginning of the interaction or with feedback. They all started with an introductory phase, in which the robot explained that they would visit a location on the tablet (e.g., the zoo), after which the robot first repeated the target words learned in the previous session (starting from session two) and continued with introducing the new target words. During this word binding phase, the tablet displayed a drawing or animation of the new target word and prompted the child to select the object or animation in Dutch (e.g., "click on the cage with one monkey"), and after the child selected this target, the tablet translated the word to English (in the example the word "one"). The robot would repeat the word and ask the child to repeat the word too. When all new target words were repeated, the child had to perform different tasks on the tablet (touching and dragging objects on the tablet screen) or had to act out target words. At the end of the session, there was a short in-game test where the child's knowledge of the target words was tested.

The recap session had a different setup because there were no new target words presented during this session. The robot first

Table 1
Target words for each domain and session.

Session	Tablet environment	Target words
Number domain		
1	Zoo	One, two, three, add, more, most
2	Bakery	Four, five, takeaway, fewer, fewest
3	Zoo	Big, small, heavy, light, high, low
Spatial domain		
4	Fruit shop	On, above, below, next to, falling
5	Forest	In front of, behind, walking, running, jumping, flying
6	Playground	Left, right, catching, throwing, sliding, climbing
Recap		
7	Photo book	Repetition of all learned words



(a) Session 1 (b) Session 6

Fig. 3. Tablet environment for session 1 and session 6.

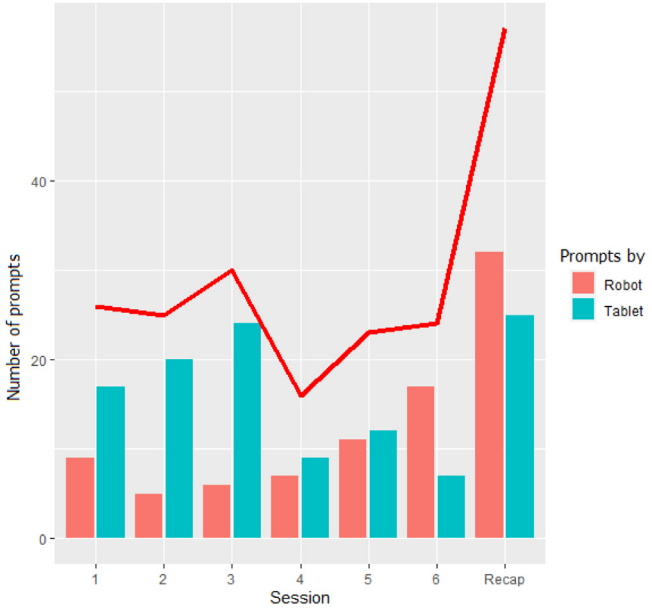


Fig. 4. Number of prompts per session. The red line shows the total amount of prompts per session. Tablet prompts contain actions such as dragging and touching objects on the screen, robot prompts contain repetition and re-enactment of the target words.

explained that it would be the last time that they were together and that they would go through all previously visited places with a photo book. Each page of the photo book contained one of the sessions with all target words. Children had to add pictures of the different target words in the photo book while repeating the words with the robot. During this session there was no in-game test in the end.

During all sessions, except during the in-game tests, the robot acted as a more knowledgeable peer that was also learning English, but provided feedback on the child's actions when needed (acting as a peer-tutor). For example, when a child was reluctant to drag an object on the tablet, the robot would first ask the child to execute the task, but after two unsuccessful attempts, the robot would perform this task for the child using a deictic gesture. The interaction was semi autonomous, the experimenter would press a button on a control panel as soon as a child had repeated the robot's speech because children's speech detection remains unreliable (Kennedy et al., 2017).

Fig. 4 shows the number of prompts that children received during each session, children had the least prompts in session 4 and the most in session 7, the recap session. The tablet and robot both prompted the child to execute tasks. Prompts by the tablet contained actions such as dragging and touching objects on the screen, prompts by the robot contained repetition and re-enactment of the target words. After successfully completing a task that was prompted by either the tablet or the robot, the robot always provided the child with feedback. This feedback could be negative feedback after an incorrect response, after which the child could try again, or positive feedback after a correct response. In other words, after each prompt, the child always received feedback from the robot.

The interaction was a one-on-one interaction, but the experimenter stayed in the same room to intervene when necessary. The duration of each session was between 15 to 25 min.

3.4. Materials

3.4.1. Measurements

Pre-test. Before the children started the seven tutoring sessions we tested their L2 knowledge of the 34 target words with an English to Dutch translation task, children's Dutch vocabulary knowledge, selective attention and their non-word repetition skills. Children were asked to translate each target word (34) from

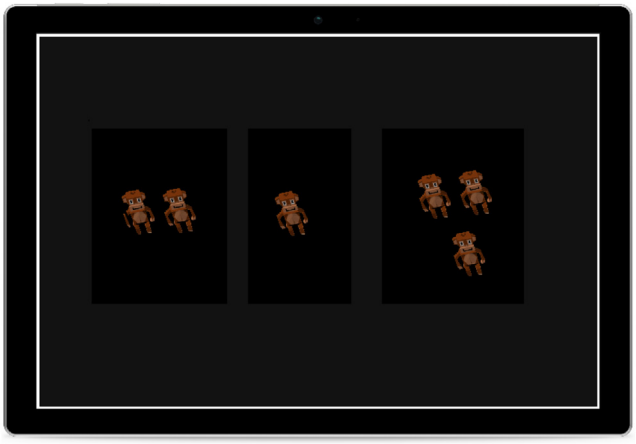


Fig. 5. In this example, in order to test the children's understanding of the word "two", the tablet asked the children to select the picture showing the two monkeys.

English to Dutch during the translation task. The target words were prerecorded by a native speaker and played through a laptop. There were two versions of the translation task, different in their word order, randomly assigned to children. Children could score 34 points on this test by providing the correct translation of the target words in Dutch. Cronbach's alpha showed that the reliability for this test was excellent, $\alpha = .96$. The main purpose of this test was to exclude children who already knew more than half of the target words before the experiment. In addition to the translation task, we measured children's Dutch vocabulary knowledge (Peabody Picture Vocabulary Test Dunn, Dunn, & Schlichting, 2005). During this task children had to select a picture out of four different pictures corresponding to the word that the experimenter said in Dutch. After making nine errors, the test stopped and the child's corresponding Dutch vocabulary level was recorded. Moreover, we measured their selective attention with a visual search task (Mulder, Hoofs, Verhagen, van der Veen, & Leseman, 2014) during which children had to search certain animals on a screen as fast as possible. Children could score a maximum score of eight. Finally, we measured children's phonological memory with a non-word repetition task (Chiat, 2015). Children had to repeat twelve not existing words in order to test their phonological memory. For each word correctly pronounced, children received one point. Cronbach's alpha showed that the reliability of this task was satisfactory, $\alpha = .76$.

We also conducted a perception questionnaire during this pre-test. However, these measurements are beyond the scope of this study. More information can be found about these measurements in van den Berghe, de Haas, Oudgenoeg-Paz, Krahmer, Verhagen, Vogt, Willemsen, de Wit, and Leseman (2021).

The total duration of the pre-test was 30–40 min. Children received a sticker for each task completed. We did not include other word knowledge tests during the pre-test to avoid the possibility that children would learn from the different tests in addition to the experimental tutoring sessions.

In-game tests. At the end of each tutoring session, children received an in-game test in which we measured their short-term retention of the target words. This in-game test was a comprehension task during which children saw three options (see Fig. 5) and the tablet asked for a certain target word. Each target word learned during that session was shown twice during the in-game test.

Post-tests. We administered two post-tests: an immediate post-test maximally two days after the recap session and a delayed post-test at least two weeks after the recap session. Both post-tests contained a translation task for all target words from Dutch to English, a translation task from English to Dutch and a comprehension task. The translation tasks were the same as the pre-test, except that children also had to translate the words from Dutch to English. Cronbach's alpha was excellent for all tests (all $\alpha \geq .94$).

The comprehension test was a picture-selection task to test the children's receptive knowledge. In this task, children were presented with a target word prerecorded by a native speaker and asked to choose which one out of three pictures or videos matched the target word ("Where do you see: *heavy*?"). Each target word was presented three times in a random order to compensate for children's guesses. Only half of the target words were included, as a test including all target words would take too long for these young children. The words included were selected in such a way that there were an equal number of words from every session. Cronbach's alpha was good, $\alpha = .84$ for the immediate post-test, for the delayed post-test, $\alpha = .87$.

3.5. Procedure

One week before the first tutoring session, children received a group introduction to familiarize themselves with the robot. During this introduction the robot explained that the children have to listen carefully and speak clearly to the robot, it also showed how it is able to move by doing a familiar dance to Dutch children and the robot shook hands with all children to reduce any anxiety that children might have towards being close to the robot.

After this introduction, each child completed the pre-test in a one-on-one setting with one of the experimenters. During the next four weeks, children (except for the children in the control condition) took part in the seven L2 tutoring sessions with the robot, each during school hours and in a one-to-one setting.

During the experimental days, the child was brought by the experimenter to a separate room with the robot and tablet to receive the session. The child was asked to sit in front of the tablet close to the robot. Before the experiment leader started the first session, he or she explained how the tablet worked and what the child was going to do with the robot. During the session, the experimenter tried to not intervene, only when the tablet game broke down or the child was reluctant to continue the session. Occasionally, the session was interrupted due to technical break downs, toilet visits or in some cases anxiety by the children. Usually the session was continued within a few minutes, but when this was not possible the experimenter returned the child to their classroom and the full session was restarted to the last point, after which the child was brought again to the robot or tablet and continued their session. Only in 9 cases, where the child did not want to proceed, the experiment was stopped and the data of these children were removed from the analyses. After the session the children were returned to their classroom, and the setup for the next child was prepared. Children received an immediate post-test within two days after the seventh session, and a delayed post-test two to five weeks after the immediate post-test. Similarly to the pre-test, the test was a one-on-one session with an experimenter. After the delayed post-test, children in the tablet-only condition were brought once more to the robot to receive one interaction with the robot in order to give them the experience of interacting with a robot.

3.6. Engagement coding

We annotated two types of engagement: task engagement and robot engagement.

Task engagement: with task engagement we measured how focused on the task the children were while executing it, whether the children were distracted and how well they responded to the questions of the tablet and the robot. Although many tasks had to be performed on the tablet (e.g., dragging animals into cages), it is important to stress that task engagement is not equivalent to tablet engagement. Task engagement contains both the engagement for the tablet as for tasks that the robot instructed. For example, each session the robot asked children to repeat words, this interaction is also part of the task. Moreover, in sessions 5 and 6, children were instructed to act out verbs such as running and jumping, which is also part of the task.

Robot engagement: this type of engagement focuses on the social aspect of engagement. For instance, how well the children imitated the robot after its gestures, how often the children looked at the robot or talked with the robot.

3.6.1. Coding scheme

Our observation scheme was adapted from the existing scheme for toddlers called ZIKO (Laevers, 2005). This observation scheme is used in preschools to observe toddlers during their activities and provides examples that raters can use to determine the child's engagement score. In the original scheme, the authors recommend to observe a minimum of 7 min per child to get a reliable engagement measure (r = .83, Laevers, 2003) (r = .89, Colpin, Laevers, & vandemeulebroecke, 2002) for a full day interaction. Because our interactions only took 15 min per session instead of a full day, we averaged the ratings of two two-minute fragments per session: two minutes in the beginning and two minutes in the end of the session. Therefore, the total average engagement rating is based on four minutes per session (see Section 3.6.4 for more information).

Similar as the ZIKO, our observation scheme consists of five levels, with five specific labels from low engagement to high engagement and four intermediate points (see Table 2). It contains example behavior that belongs to a certain engagement level. The scheme is organized in such a way that children who did not show any interest, or were continuously talking to the experiment leader were rated with a low level of engagement and children who were continuously working and were completely absorbed were rated with a high level of engagement. Children who executed everything but did not show any interest fell in between, received a medium engagement score. See Table 2 for a few example behaviors for each level. The full engagement scheme contains more examples and can be found in the Appendix and on Github.[6]

The same levels were used for both engagement types (task and robot). However, the examples and explanations of the levels were adapted for the specific engagement type. For *task engagement*, we used the same examples as the ZIKO scheme, however, we added a few examples that were specific for our interaction (e.g., The child meaninglessly touches the tablet (low engagement), looks the whole time at the task environment or robot (high engagement)). *Robot engagement* used similar examples as task engagement, however they were changed into social interaction moments (e.g., "No signs of interest" was changed into "No signs of interest in the robot" (low engagement) and "enjoys being so driven" was changed into "enjoys working with the robot" (high engagement)). Furthermore, we added specific examples for robot engagement such as, "ignores the robot fully" (low robot engagement), "purposelessly touching the robot" (average engagement), "talks to the robot", "there is joint attention" (high engagement).

3.6.2. Annotation software

We developed our own annotation software program for the raters (see Fig. 6). The program showed a front view and a side view of the two-minute fragment and contained a short table with the examples of the different engagement levels. Raters could check the example behaviors in a table while watching the fragment. They were not allowed to stop the video during the two minutes and had to wait until it was finished. However, they could already write comments to help forming their rating about the video. The tool automatically saved all ratings.

3.6.3. Engagement coding

The first author together with nine student assistants annotated the data. The nine student assistants received a group training from the first author. This training took one full-day and raters practiced with ten different videos. After the training all raters received a summary of the training, the annotation scheme and the annotation program. During the annotation period, there were biweekly sessions during which difficult video fragments were discussed and during which the group decided on a final rating for these specific fragments. Part of the videos was double rated by different pairs of raters and their inter-rater agreement was considered moderate using the intraclass correlation coefficient (ICC = .72, 95%CI[.70, .74] (Koo & Li, 2016)). While this score is lower than reported for the original scheme (ICC = .83), it is very consistent with other studies in the field of child-robot interaction using this method, such as (de Wit, Brandse, Krahmer, & Vogt, 2020), which reports a range of ICC scores from .45 to .83, and van Minkelen et al. (2020), with a range of .6 to .89. Therefore, we consider the score for this study sufficient for further analysis. We used the raters' weighted average during our analyses.

3.6.4. Videos fragments

The videos in this data set were cut into two two-minute fragments: one in the beginning of the video and one in the end of the video. These fragments were chosen to include multiple interactions between the robot and child. For example, the first fragment always started at the beginning of the concept binding phase, and therefore included not only the first introduction to words, but also the application of the target words in other settings such as dragging animals into the cage. The second fragment was timed in such a way that it showed the end of the interaction, before the in-game tests would start. The mean of these two fragments resulted in an average engagement and was used for the analyzes on engagement. We excluded interactions during which children had a break, for instance when they had to go to the toilet or a crash occurred (9%) because an interruption could have influenced their engagement. Some videos were lost during the experiments (2%). Furthermore, there were videos that were not suitable for analyses, for instance the lighting was too dark or the video was corrupted or the recording started after beginning the experiment, which made it too difficult to find the same fragments for each child (16%). Finally, some videos had the wrong naming or the video stopped halfway (2%). This resulted in a data set containing 817 unique videos with 1635 different fragments, which is 73% of all possible data. For these fragments we annotated task engagement. Robot engagement was only annotated for the robot conditions and resulted in 537 sessions and 1074 different fragments.

[6] See https://github.com/l2tor/codingscheme

Table 2
A part of the engagement coding scheme as used in this experiment. The full coding scheme can be found on Github (see footnote 4) and in the Appendix.

Level	Engagement	Task engagement examples	Robot engagement examples
1	Very low	*The child shows virtually no activity:* - no concentration, - only ticking on the screen to continue the game, - only concerned with the experiment leader and not with the task.	*The child shows virtually no interaction with the robot:* - ignores the robot completely, - has a closed body position towards the robot, - no signs of interest in the robot.
2	Low	*The child shows some activity, but is regularly interrupted:* - limited concentration, - fidgeting, - easily distracted.	*The child shows some robot interaction, but is regularly interrupted:* - looking away, - limited looking at the robot, - easily distracted.
3	Medium	*There is activity all the time, but not really focused:* - has limited motivation, - does not feel challenged, only uses his capacities in moderation, - most tasks are performed.	*The child is active with the robot all the time, but not really focused:* - has an open body attitude towards the robot, - aimlessly touching the robot, - the child is not absorbed in his game with the robot.
4	High	*The child is mostly engaged with the task:* - the child is totally absorbed in his game, - the child feels challenged, - there is a certain drive.	*The child is mostly engaged with the robot:* - the child is absorbed in his game with the robot, - there are usually signs of joint-attention, - there is usually concentration but sometimes the attention drops.
5	Very high	*The child is completely absorbed in his activity with the task:* - is continuously concentrated, - forgets about the time, is very motivated, - enjoys being so engaged.	*The child is completely absorbed in his activity with the robot:* - there are signs of joint attention, - talks to the robot, looks at the robot, - enjoys being so engaged with the robot.

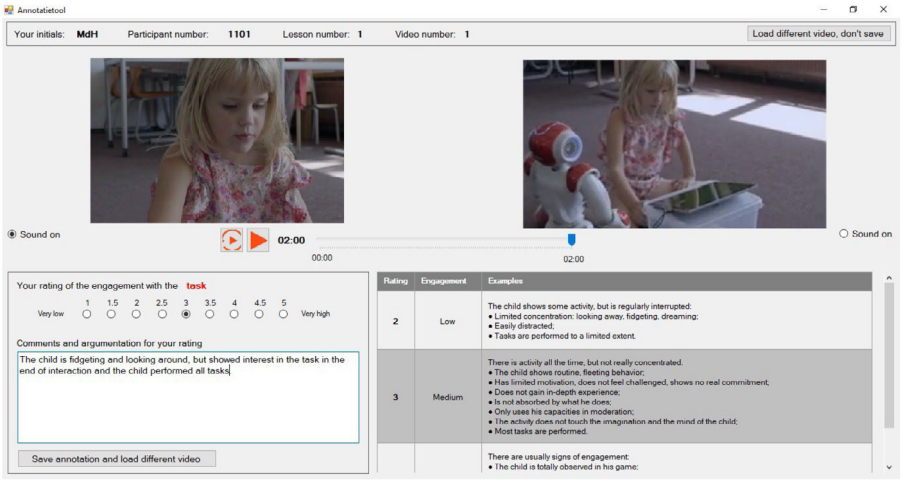


Fig. 6. Annotation tool for engagement raters. Note that this rating is an example and is not taken from our dataset.

3.7. Analyses

Task engagement was rated for the three tutoring session conditions and not for the control condition, and robot engagement was only rated for the tutoring sessions with the robot present.

First, we investigated whether children's task engagement and robot engagement changed over time and conditions. We used a mixed design ANOVA to compare children's task engagement and robot engagement over the different sessions within the different conditions. Because of missing values in the data file, it was not possible to perform pair-wise comparisons between all sessions. Moreover, the relation between sessions and engagement did not seem linear but quadratic, therefore, for the post-hoc analysis, we conducted a quadratic regression analysis.

We also explored effects of gender and age on both engagement types, using a t-test to compare the different genders and a linear regression analysis for age. Second, we used Pearson's correlations to investigate how task engagement, robot engagement were related with children's knowledge of L2 words. We correlated the average of children's task engagement, the average of children's robot engagement and scores on immediate post-test, delayed post-test. Finally, we did an exploratory analysis whether children's selective attention, children's general Dutch knowledge and children's non-word repetition were correlated with children's task and robot engagement.

4. Results

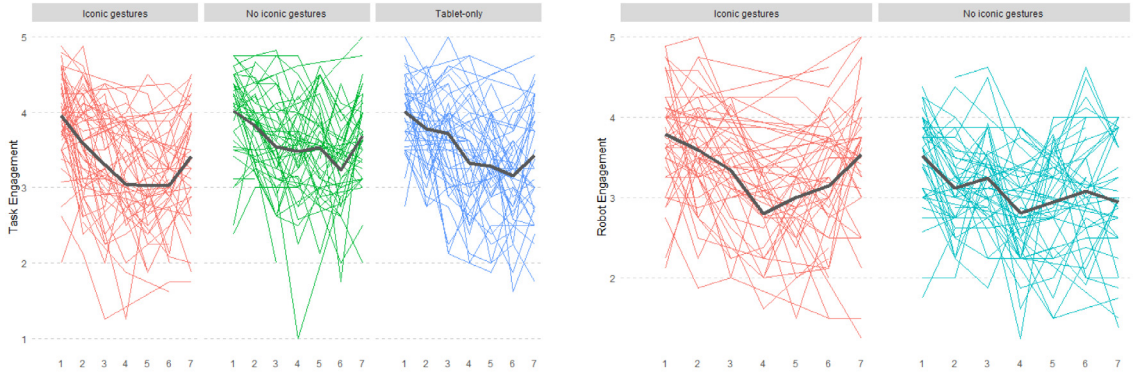
We investigated the relation between task engagement and robot engagement using Pearson's correlation. Task engagement and robot engagement were moderately correlated, $r(525) = 0.52$, $p < .001$. The relation was positive, suggesting that children who were more engaged with the task were also more engaged with the robot. However, the correlation is not very high, which

Table 3
An overview of our findings.

	Task engagement	Robot engagement
Conditions	No significant difference	Robot engagement was higher for a robot using iconic gestures than without iconic gestures
Sessions	Quadratic relation	No difference
Word knowledge	Positive relation	Retention word knowledge correlates with robot engagement

Table 4
Task engagement and robot engagement scores for each session (SD).

Engagement	Total	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6	Recap
Task								
Iconic gestures	3.41 (0.75)	3.96 (0.58)	3.59 (0.66)	3.31 (0.74)	3.05 (0.77)	3.03 (0.70)	3.03 (0.63)	3.41 (0.73)
No iconic gest.	3.64 (0.64)	4.02 (0.56)	3.82 (0.45)	3.54 (0.66)	3.48 (0.72)	3.54 (0.67)	3.23 (0.59)	3.68 (0.61)
Tablet-only	3.54 (0.70)	4.01 (0.46)	3.78 (0.50)	3.72 (0.70)	3.32 (0.74)	3.28 (0.64)	3.16 (0.65)	3.43 (0.75)
Robot								
Iconic gestures	3.39 (0.76)	3.78 (0.63)	3.59 (0.66)	3.35 (0.67)	2.80 (0.63)	3.00 (0.60)	3.14 (0.77)	3.54 (0.85)
No iconic gest.	3.11 (0.67)	3.52 (0.55)	3.12 (0.55)	3.25 (0.58)	2.81 (0.64)	2.95 (0.70)	3.08 (0.74)	2.94 (0.75)



(a) Task engagement (b) Robot engagement

Fig. 7. The individual children's engagement ratings over time and per condition. The black line shows the average engagement during each session.

confirms that there is a difference between the two engagement types and shows that we measured two related, yet distinct aspects of engagement in the interaction. An overview of our main findings can be found in Table 3. The details of these findings can be found in the following sections.

4.1. Engagement over time and conditions

4.1.1. Task engagement

Table 4 and Fig. 7a show children's task engagement over time for the three conditions. Each line in Fig. 7a represents the task engagement of an individual child (thus highlighting the individual differences) and the black lines show the averages. The figure shows that task engagement tends to drop over time, however the task engagement increases again during the recap session.

We conducted a mixed design ANOVA with the children's task engagement scores as dependent variable, and with sessions as within factor and condition as between factor to investigate the relation of task engagement and time in the different conditions. There was a main effect of session on task engagement ($F(6, 138) = 7.98, p < .001, \eta^2 = .18$). However, there was no significant difference in task engagement between conditions ($F(2, 23) = 1.43, p = .26, \eta^2 = .05$). Children were similarly task-engaged in all conditions. Nor was there a significant interaction effect between task engagement over sessions and the different conditions ($F(12, 138) = 0.80, p = .65, \eta^2 = .04$).

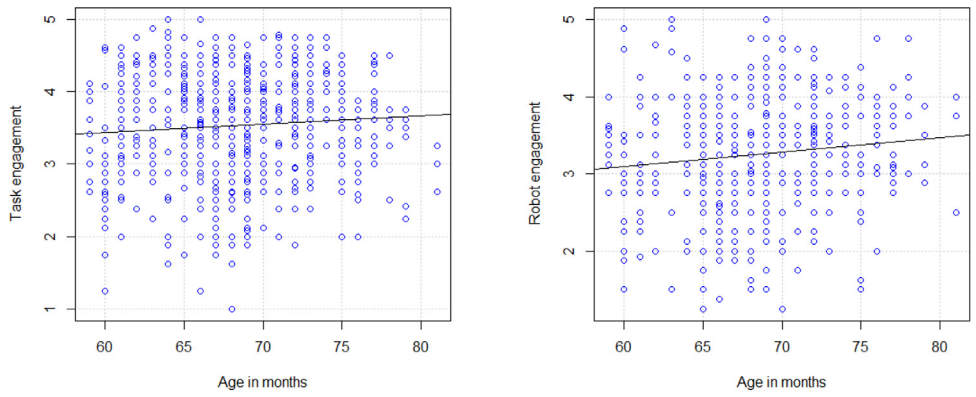
Furthermore, in order to explore the effect of session on task engagement, we conducted a quadratic regression model. This model showed that session significantly predicted task engagement: task engagement $= 4.44 - 0.46$ * session number - 0.04 * session number^2.

Finally, we checked for demographic variables on the full data set. There was no significant effect for gender ($t(807) = -1.44, p = .15$). Boys ($M = 3.50, SD = 0.71$) and girls ($M = 3.57, SD = 0.69$) did not differ in their task-engagement scores. Moreover, a linear regression analysis showed a weak interaction effect of age on task engagement. Age significantly predicted task engagement; ($F(1, 807) = 4.70, p = .03, R^2 = .006$). Children's predicted task engagement is equal to $2.72+0.08$ *(age in months). Fig. 8a shows that a younger age was associated with a lower task engagement, however the regression is exceptionally weak.

4.1.2. Robot engagement

Fig. 7b shows that there are substantial individual differences between children and that the children's overall robot engagement decreased over time for each condition, similarly as for task engagement.

We used a mixed design ANOVA with robot engagement as the dependent variable and sessions as within factor and condition as between factor to investigate the relation of robot engagement and time in the different conditions. Unlike task engagement, there was a significant effect of condition for robot engagement



(a) Task engagement (b) Robot engagement

Fig. 8. Age plotted against (a) task engagement and (b) robot engagement.

Table 5
An overview of children's word knowledge scores.
Source: Table adapted from (Vogt et al., 2019).

Condition/Test	Pre-test	Immediate post-test	Delayed post-test
Iconic gestures			
Trans (En-Du)	3.31 (3.09)	7.41 (5.17)	8.10 (5.06)
Trans (Du-En)		6.00 (4.23)	6.45 (4.62)
Comprehension		29.47 (5.85)	30.43 (6.22)
No iconic gestures			
Trans (En-Du)	3.47 (3.19)	7.69 (4.92)	7.88 (4.79)
Trans (Du-En)		6.43 (4.20)	6.43 (4.65)
Comprehension		29.39 (6.08)	29.75 (6.44)
Tablet-only			
Trans (En-Du)	4.04 (2.76)	7.96 (4.63)	8.63 (4.62)
Trans (Du-En)		6.57 (4.01)	6.67 (4.20)
Comprehension		29.73 (6.27)	30.25 (6.58)

Note: All scores indicate the average number of words correctly translated or comprehended (standard deviation within brackets). Minimum scores are 0, maximum scores are 34 for translation and 54 for comprehension. For comprehension, chance level is 18.

($F(1, 13) = 6.74, p = .02, \eta^2 = .15$). Children's robot engagement was higher when interacting with the robot with iconic gestures ($M = 3.39, SD = 0.76$) than with the robot without iconic gestures ($M = 3.11, SD = 0.67$). We found no significant effect over sessions on robot engagement ($F(6, 78) = 1.50, p = .19, \eta^2 = .07$), Moreover, there was no significant interaction effect of robot engagement over sessions in the different conditions ($F(6, 78) = 1.39, p = .23, \eta^2 = .07$).

Similarly as task engagement, there was no effect of gender on robot engagement. Boys ($M = 3.27, SD = 0.71$) and girls ($M = 3.22, SD = 0.75$) did not differ in their robot engagement scores ($t(527) = 0.86, p = .39$).

Finally, again similar as task engagement, a weak interaction effect of age was found on robot engagement, a linear regression analysis showed that age significantly predicted robot engagement; ($F(1, 527) = 6.98, p = .009, R^2 = .013$). Children's predicted robot engagement is equal to $1.99 + 0.11$ *(age in months). Fig. 8b shows that a younger age was associated with a lower robot engagement, however the explained variance is very small.

4.2. Relation engagement and word knowledge

Table 5 displays children's word knowledge scores on the pre-test, immediate post-test and delayed post-test. To investigate whether there is a relation between the performance of the children and their engagement, we calculated correlations between their word knowledge scores and task engagement and robot engagement.

As Table 6 shows, there were many weak, yet significant, correlations between the children's learning performances and their engagement with both the task and the robot. Children's task engagement correlates significantly with all pre-test and post-test word knowledge scores. Task engagement also significantly correlates with children's selective attention and non word repetition. Robot engagement only correlates with the pre-test, the immediate translation task from Dutch to English and all delayed post-tests, where the correlation is slightly higher for the two translation tasks. In contrast to task engagement, selective attention is not correlated to children's robot engagement. Their non-word repetition is negatively correlated, suggesting children are less robot-engaged when children are better in pronunciation of non words and vice versa.

4.3. Relation engagement and prompts

We also explored the relation between the prompts children received and children's engagement. We calculated the correlation between children's engagement and the number of times a session required children to interact with the tablet (touch an object or move an object) or with the robot (repeat the robot's speech or repeat the robot's gesture) as shown in Fig. 4. As Table 7 shows, prompts by the tablet showed a positive relation with children's task engagement and robot engagement, prompts by the robot showed a negative relation with children's task engagement and robot engagement. In other words, children's task and robot engagement increased when children had to interact more

Table 6
Correlations between children's English word knowledge and their engagement.

		Task Engagement	Robot Engagement
Pre-test	Translation (En-Du)	.08*	.09*
Immediate Post-test	Translation (En-Du)	.09*	.08
	Translation (Du-En)	.14***	.10*
	Comprehension	.13***	.08
Delayed Post-test	Translation (En-Du)	.13***	.15***
	Translation (Du-En)	.12***	.15***
	Comprehension	.12***	.09*
Selective attention		.17***	−.03
Non word repetition		.10**	−.10*
Dutch receptive vocab		.04	−.03

Note: * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 7
Correlations between the prompts during the game and children's engagement.

Prompts by	Task engagement	Robot engagement
Tablet	.21***	.18***
Robot	−.11**	−.04
Total	.04	.07

Note: * $p < .05$, ** $p < .01$, ***$p < .001$.

often with the tablet and vice versa (task: $r(807) = .21, p < .001$, robot: $r(527) = .18, p < .001$). In contrast, when children had to interact more with the robot, children's task engagement significantly decreased and vice versa ($r(807) = -.11, p = .002$). Note that these prompts only refers to the interaction required by the task (manipulating objects on the tablet, required verbal and non-verbal behavior towards the robot), and not to the unscheduled interaction between the game and child (e.g., the robot's feedback).

5. Discussion

The aim of the present study was to examine how children's task engagement and robot engagement developed over time in a long-term child-robot interaction for second-language tutoring. More specifically, we compared children's task engagement when interacting with (a) a robot using iconic gestures and a tablet, (b) a robot without iconic gestures and a tablet, and (c) only with a tablet. Furthermore, we compared children's robot engagement with (a) a robot using iconic gestures and (b) a robot without iconic gestures. Lastly, we compared children's second-language word knowledge with their task engagement and robot engagement.

Although task engagement and robot engagement were only moderately correlated, the two are inherently connected and show the same trends (Oertel et al., 2020). There were large individual differences between children over sessions but overall, both task engagement and robot engagement decreased over time and increased again with the seventh session. The decreasing pattern is weak due to the high variance in engagement. Both engagements seemed to fluctuate less after the third session, that might indicate that the novelty effect plays a smaller role after the third session, which also has been reported by Salter et al. (2004). These findings suggest that, overall, children were very excited to interact with the robot and tablet in the first few sessions, but after some sessions, the robot, tablet and tasks were not as new and exciting anymore and children returned to their normal, less engaged behavior.

5.1. Task engagement

We investigated children's task engagement during all seven sessions with the robot. Overall, children's task engagement decreased over sessions, in line with other long-term studies (Kanda et al., 2004; Leite, Castellano, Pereira, Martinho, & Paiva, 2014; Serholt & Barendregt, 2016).

Contrary to our expectation (H1a), children were not more task-engaged in the robot conditions than in the tablet-only condition nor did we find an effect of condition over time (H1b). A possible explanation for this finding is that there were large individual differences between children in task engagement over sessions, something that other studies also reported (de Haas et al., 2020). These differences can explain why we did not find large statistical differences between conditions, because the large individual differences would be a factor for high variance in the data. The study by Van den Berghe, Oudgenoeg-Paz, Verhagen, Brouwer, De Haas, De Wit, Willemsen, Vogt, Krahmer, and Leseman (2021) discussed the individual differences of this study in more detail.

This finding can likewise be explained in an alternative manner. In all conditions, the task remained constant, and only the presence of the robot and the use of gestures varied. Consequently, in retrospect, it may be not so unexpected that children's task engagement did not differ across the three conditions, as their task engagement relates to the task itself, which remained the same.

We did find an overall effect of time on children's task engagement: children's task engagement decreased during the sessions, and increased during the last session (the recap session). This increase is likely due to the nature of the last session, which was a recap session and different than the other sessions. During the recap session, children had to speak to the robot, click on the screen and move all the different target words they had learned during the tutoring sessions. This created a highly interactive session and suggests a link between children's task engagement and interaction with the tablet. The difference between the other sessions and recap may also have resulted in a re-introduction of the novelty effect and thus increased children's task engagement. This same finding has been shown in earlier experiments (Davison et al., 2020; Tanaka et al., 2007). Likewise, it is also possible that because session 7 was a recap session and the children recognized the words, their task engagement increased because they recognized the target words. Each session (except for the first) started with a small recap and in some cases, children expressed that they recognized the words but were not sure about the meaning anymore. During the recap session, children could chose the meaning of the word from a few options (receptive knowledge instead of active knowledge), and the target words were more easily recognized.

When comparing task engagement and the prompts in the sessions, children's task engagement was, as expected, positively related to the tablet's prompts (e.g., dragging an object on the screen, selecting an object on the screen). Interestingly, there was a negative relation between the prompts by the robot (speech and re-enactment of gestures) and children's task engagement. This was unexpected because these prompts by the robot were

also part of the task. This negative relationship may possibly be explained by the fact that these interaction moments with the robot may have created anxiety for shy children because they had to talk to the robot in an unfamiliar language and, as a result, their discomfort made them less engaged with the task. This would also explain why the correlation was weak, not all children felt uncomfortable speaking a second language. This also accords with the positive correlation between children's non-word repetition and children's task engagement. Children who repeated more words correctly (and possibly more confidently) during the pre-test, also scored higher on task engagement and children who scored lower on the non-word repetition task, and therefore were less likely to actively repeat the robot during the interaction, scored lower on task engagement.

To get an additional idea of other aspects that could affect children's task engagement, we performed exploratory analyses on age and gender. These analyses showed that age was related to task engagement: younger children were less task-engaged than older children. The effect was small, which is likely due to the fact that the age variation in our experiment was also relatively small because all children were in the same year at school. There are at least two possible explanations for the relation between age and task engagement. Younger children tend to have a shorter attention span than older children, and were therefore more likely to get distracted during the task and become less task-engaged (Betts, Mckay, Maruff, & Anderson, 2006). This is confirmed by the correlation between children's selective attention and task engagement. It seems that children who have a larger selective attention and can therefore focus longer on one particular task, are more task-engaged during the experiment. Another possible explanation for this is that it is harder to observe whether or not younger children are engaged and that older children demonstrated the typical behaviors related to engagement more frequently in the way we expected them to. We did not find differences between girls and boys, overall both genders showed similar levels of task engagement.

5.2. Robot engagement

Unlike task engagement, children were more robot-engaged with a robot that used iconic gestures than with a robot without iconic gestures (confirming H2a). This is line with a study by de Wit et al. (2020) who found that 5-year-old children were more robot-engaged with a robot using gestures than a robot using no gestures, although their experiment only contained one session. The iconic gestures by the robot contributed to a higher robot engagement, which can be explained by the fact that a robot that moves physically, attracts more attention and appears more active, thereby stimulating the child to remain robot-engaged. In the condition without gestures, the robot was less active and therefore children were less attentive and engaged towards the robot.

Moreover, unlike task engagement, children's robot engagement did not decrease significantly over the different sessions (H2b). When looking more closely at these different sessions, some observations can be made. Robot engagement dropped most during session four. This can be plausibly explained by the number of prompts during sessions. Session four had the fewest prompts of all sessions and could therefore have resulted in the lowest robot engagement.

Robot engagement increased in the recap session, however only in the iconic gestures condition and not in the without iconic gesture condition. This observed increase could be attributed by the variety of iconic gestures during the recap session in contrast to the repetitiveness of the robot gestures in the other sessions. In each session, there were at least five target words that used the same iconic gesture. In the recap session, all 35 target words were repeated twice, and therefore the robot showed a larger variety of gestures that might have sparked children's robot engagement. Future studies can investigate whether a variation of gestures during the sessions itself will sustain children's robot engagement over time more than repeating the same gesture.

We found a positive correlation between robot engagement and the prompts by the tablet but surprisingly, no significant correlation between robot engagement and the robot's prompts. It is difficult to explain these findings, but it is important to note that these prompts only focused on the interaction moments (i.e. when the children had to respond in one way or another) as implemented in the game. For instance, this positive correlation between robot engagement and the tablet prompts may actually be due to the robot's feedback. During the prompts by the tablet, after children touched or dragged an object on the screen, the robot would provide the child with positive feedback. Thus, these positive feedback was not part of the prompts by the robot, but always followed the prompts by the tablet. Arguably, this positive feedback by the robot increased children's robot engagement. This also accords with observations in de Haas et al. (2020), who showed that the type of robotic feedback has influence on children's task engagement and robot engagement.

In addition, we noticed that children spontaneously re-enacted the gestures or were spontaneously talking with the robot about the game or other events. It is likely that these spontaneous moments with the robot increased children's robot engagement more than by the game initiated prompts. This is in line with Ahmad et al. (2017), who found that not game adaptation, but emotion-based adaptations during child–robot interactions sustained robot engagement over time.

5.3. Relation engagement with word knowledge

Finally, we investigated the relation between children's engagement and their word knowledge to see to what extent engagement relates to learning outcomes. We found a weak but significant correlation between task engagement and children's word knowledge (confirming H3a). This confirms previous studies that describe that there is a link between children's word knowledge and their engagement (e.g., Blumenfeld, Kempler, & Krajcik, 2006; Linnenbrink & Pintrich, 2003). The effect seems to be stronger for children's delayed word knowledge, which might be due to the fact that children who are more task-engaged, remember the task more vividly including the target words and therefore retain more word knowledge over time.

Unlike we expected, children's robot engagement did not negatively influence word knowledge (H3b). In fact, children's robot engagement was, similar as task engagement, positively correlated with the pre-test, the immediate translation task from Dutch to English and all delayed post-tests. It has been suggested that children who are more robot-engaged get distracted from the task and learn less (Kennedy et al., 2015). This does not appear to be the case in our experiment. Our findings show that there is a positive link between children's robot engagement and learning gain. The effect seems to be stronger for children's delayed word knowledge, children who were more robot-engaged might recall the interaction more often and therefore remember more words which results in a higher score on the delayed post-test.

However, these results must be interpreted with caution because the correlations were weak, though statistically significant. Moreover, the results do not show a causal relation between engagement and word knowledge. The design of the study did not allow to investigate a causal relation between these two factors. It is therefore not possible to determine whether task and robot engagement increased children's L2 word knowledge, or whether children's L2 word knowledge increased robot engagement. A further study with more focus on the direction of this effect is therefore suggested.

5.4. Limitations and strengths

Our study has multiple limitations. First, our interactions between robot and child were rather fixed and did not include any adaptation when children became disengaged. A change of the robot's behavior could possibly have increased children's robot engagement and made them more engaged with the game again (Ahmad et al., 2019; Tanaka et al., 2007). However, for experimental soundness, our design allowed us to compare children in the four different conditions without any other interaction differences between them. Future studies can take our findings into account and focus on possible ways of changing the robot's behavior while still keeping the children focused on the task. Second, we could only investigate correlations between task engagement, robot engagement and word knowledge and no causal relations. Therefore, we cannot determine whether children's L2 word knowledge will become higher with a higher task or robot engagement or vice versa. Future research is needed to test whether an engaging task will increase children's L2 word knowledge or whether an increase in L2 word knowledge will also increase children's engagement. Third, because we focused on a specific age group, we cannot generalize our results to other ages. Our findings do suggest that older five years old children are more task and robot-engaged than younger five years old children, which leads us to believe that with other age groups, older children will be more engaged than younger children. Fourth, the robot's use of iconic gestures lengthened the experiment, which can potentially decrease children's engagement given that usually longer sessions demand longer attention span and therefore lead to a decrease in engagement. However, in our experiment it seems unlikely that the longer duration had a negative effect on children's engagement, because we found that children were *more* robot-engaged when interacting with a robot using iconic gestures and we found no difference between children's task engagement in the different conditions. Therefore, we do not think the length had a large negative effect on children's engagement.

Our study also has several strengths. It is one of the first studies to investigate task engagement and robot engagement over a long-term tutoring interaction and the relation to children's word knowledge. Moreover, we included a large sample of young children, preregistered the study before the experiment and made the source code publicly available. Lastly, we applied a new coding scheme, based on a validated approach, which is made publicly available and can be used by other researchers as a structured way of measuring task and robot engagement.

6. Conclusion

In this study, we present one of the first large-scale studies that investigated children's task engagement and robot engagement during multiple robot sessions and whether these two types of engagement were related to children's second-language word knowledge. We were particularly interested in whether children's task engagement and robot engagement differed when children interacted with a tablet and robot using iconic gestures, a tablet and a robot that did not use iconic gestures, or only a tablet. Our findings show that the robot's iconic gestures did not have an effect on children's task engagement over time, however it did have an effect on children's robot engagement. Children were more robot-engaged when a robot used iconic gestures than without iconic gestures. Moreover, children's task engagement and robot engagement were both positively correlated with children's word retention. Children who were more task-engaged or more robot-engaged knew more words two weeks after the tutoring sessions. Our findings have provided a deeper insight into the influence of the robot's gestures on children's task and robot engagement and the importance of both engagement types on children's word knowledge. As a next step, adding to the results in this study, further research is needed in order to improve the understanding of the influence of various aspects of the robot's behavior, such as robotic feedback or variation of gestures, on children's task engagement and robot engagement in long-term child-robot interactions.

Selection and Participation

The participants in our study were five-year-old children attending one of nine different primary schools in the Netherlands. The study took place at the children's school, in a quiet area designated to the experimental setup. Data related to the study were collected after approval from Ethical Review Board of the Faculty of Social Sciences of Utrecht University, following all the regulations and recommendations for research with children. Information about the study was distributed via the schools to all the legal guardians of children in the suitable ages. The information included information about the goal of the study and what was required from parents and children if they agree to participate. Legal guardians then informed the teachers if they agreed to participate in the study with their child. Only after legal guardians handed in a signed informed consent form to the teacher, did the teacher share their details with the researchers. Children were informed about the data collection process and their participation in the study was completely voluntary. In addition, children and parents were able to withdraw their consent for the data collection at any time without affecting their participation in the activity. Data was stored on secured servers of Utrecht University and only the researchers had access to any personal data. Following data collection, the data were fully anonymized.

CRediT authorship contribution statement

Mirjam de Haas: Conception and design, Software, Data collection, Analysis, Interpretation of the data, Drafting the article and revising it critically. **Paul Vogt:** Funding acquisition, Conception and design, Revising the article critically. **Rianne van den Berghe:** Conception and design, Data collection, Revising the article critically. **Paul Leseman:** Funding acquisition, Conception and design. **Ora Oudgenoeg-Paz:** Conception and design, Data collection, Revising the article critically. **Bram Willemsen:** Conception and design, Data collection. **Jan de Wit:** Conception and design, Software, Data collection, Revising the article critically. **Emiel Krahmer:** Funding acquisition, Conception and design, Revising the article critically.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

All authors approved the final version of the manuscript

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijcci.2022.100501.

References

Ahmad, M. I., Mubin, O., & Orlando, J. (2017). Adaptive social robot for sustaining social engagement during long-term children–robot interaction. *Int. J. Hum.-Comput. Interact.*, *33*(12), 943–962. http://dx.doi.org/10.1080/10447318.2017.1300750.

Ahmad, M. I., 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. *Adaptive Behavior*, *27*(4), 243–266. http://dx.doi.org/10.1177/1059712319844182.

Alimardani, M., & Hiraki, K. (2020). Passive brain-computer interfaces for enhanced human-robot interaction. *Front. Robotics AI*, *7*, 125. http://dx.doi.org/10.3389/frobt.2020.00125, URL https://www.frontiersin.org/article/10.3389/frobt.2020.00125.

Alves-Oliveira, P., Sequeira, P., Melo, F. S., Castellano, G., & Paiva, A. (2019). Empathic robot for group learning: A field study. *J. Hum.-Robot Interact.*, *8*(1), http://dx.doi.org/10.1145/3300188.

Arnott, L., Grogan, D., & Duncan, P. (2016). Lessons from using ipads to understand young children's creativity. *Contemp. Issues Early Child.*, *17*(2), 157–173. http://dx.doi.org/10.1177/1463949116633347, http://arxiv.org/abs/DOI: 10.1177/1463949116633347[arXiv:DOI: 10.1177/1463949116633347].

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*, *3*(21), eaat5954. http://dx.doi.org/10.1126/scirobotics.aat5954.

Belpaeme, T., Vogt, P., van den Berghe, R., Bergmann, K., Göksun, T., de Haas, M., et al. (2018). Guidelines for designing social robots as second language tutors. *Int. J. Soc. Robotics*, 1–17.

van den Berghe, R., de Haas, M., Oudgenoeg-Paz, O., Krahmer, E., Verhagen, J., Vogt, P., et al. (2021). A toy or a friend? children's anthropomorphic beliefs about robots and how these relate to second-language word learning. *Journal of Computer Assisted Learning*, *37*(2), 396–410. http://dx.doi.org/10.1111/jcal.12497.

Van den Berghe, R., Oudgenoeg-Paz, O., Verhagen, J., Brouwer, S., De Haas, M., De Wit, J., et al. (2021). Individual differences in children's (language) learning skills moderate effects of robot-assisted second language learning. *Frontiers in Robotics and AI*, 259.

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., & Leseman, P. (2019). Social robots for language learning: A review. *Rev. Educ. Res.*, *89*(2), 259–295. http://dx.doi.org/10.3102/0034654318821286, http://arxiv.org/abs/DOI: 10.3102/0034654318821286[arXiv:DOI: 10.3102/0034654318821286].

Betts, J., Mckay, J., Maruff, P., & Anderson, V. (2006). The development of sustained attention in children: The effect of age and task load. *Child Neuropsychol.*, *12*(3), 205–221. http://dx.doi.org/10.1080/09297040500488522, http://arxiv.org/abs/DOI: 10.1080/09297040500488522[arXiv:DOI: 10.1080/09297040500488522], PMID: 16837396.

Blumenfeld, P. C., Kempler, T. M., & Krajcik, J. S. (2006). Motivation and cognitive engagement in learning environments. In R. K. Sawyer (Ed.), *The Cambridge Handbook of: The Learning Sciences* (pp. 475–488). New York, NY, US: Cambridge University Press.

Chaspari, T., Al Moubayed, S., & Fain Lehman, J. (2015). Exploring children's verbal and acoustic synchrony: Towards promoting engagement in speech-controlled robot-companion games. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy and InfLuence* (pp. 21–24).

Chiat, S. (2015). Non-word repetition. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Methods for Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment* (pp. 227–250). Bristol: Multilingualism Matters.

Christenson, S. L., Wylie, C., & Reschly, A. L. (2012). *Handbook of Research on Student Engagement* (pp. 1–840). New York, NY, US: Springer Science & Business Media, http://dx.doi.org/10.1007/978-1-4614-2018-7.

Chung, E. Y.-h. (2019). Robotic intervention program for enhancement of social engagement among children with autism spectrum disorder. *J. Dev. Phys. Disabil.*, *31*(4), 419–434.

Colpin, H., Laevers, F., & vandemeulebroecke, G. (2002). Ontwikkeling van een evaluatie-instrument voor de opvang in en door vlaamse basisscholen met het oog op het verkrijgen van een kwaliteitslabel: onderzoeksinstrumenten. *Onderzoeksrapport*, *35*.

Davison, D. P., Wijnen, F. M., Charisi, V., van der Meij, J., Evers, V., & Reidsma, D. (2020). Working with a social robot in school: A long-term real-world unsupervised deployment. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 63–72). New York, NY, USA: Association for Computing Machinery.

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 *HRI '20, Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 359–367). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3319502.3374815.

de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., et al. (2018). The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. In *HRI '18, Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 50–58). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3171221.3171277.

Díaz, M., Nuño, N., Saez-Pons, J., Pardo, D. E., & Angulo, C. (2011). Building up child-robot relationship for therapeutic purposes: From initial attraction towards long-term social engagement. In *Face and Gesture 2011* (pp. 927–932). IEEE.

Dunn, L. M., Dunn, L. M., & Schlichting, L. (2005). *Peabody Picture Vocabulary Test-III-NL*. Amsterdam: Pearson.

Filippini, C., Spadolini, E., Cardone, D., Bianchi, D., Preziuso, M., Sciarretta, C., et al. (2020). Facilitating the child–robot interaction by endowing the robot with the capability of understanding the child engagement: The case of mio amico robot. *Int. J. Soc. Robotics*, 1–13.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Rev. Educ. Res.*, *74*(1), 59–109. http://dx.doi.org/10.3102/00346543074001059, http://arxiv.org/abs/DOI: 10.3102/00346543074001059[arXiv:DOI: 10.3102/00346543074001059].

de Haas, M., Krahmer, E., & Vogt, P. (2020). The effects of feedback on children's engagement and learning outcomes in robot-assisted second language learning. *Front. Robotics AI*, *7*, 101. http://dx.doi.org/10.3389/frobt.2020.00101, URL https://www.frontiersin.org/article/10.3389/frobt.2020.00101.

Heath, S., Durantin, G., Boden, M., Hensby, K., Taufatofua, J., Olsson, O., et al. (2017). Spatiotemporal aspects of engagement during dialogic storytelling child–robot interaction. *Front. Robotics AI*, *4*, 27.

Ishii, R., & Nakano, Y. I. (2010). An empirical study of eye-gaze behaviors: Towards the estimation of conversational engagement in human-agent communication. In *Proceedings of the 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction* (pp. 33–40).

Jacq, A., Lemaignan, S., Garcia, F., Dillenbourg, P., & Paiva, A. (2016). Building successful long child-robot interactions in a learning context. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 239–246). IEEE, http://dx.doi.org/10.1109/HRI.2016.7451758.

Jang, H. (2008). Supporting students' motivation, engagement, and learning during an uninteresting activity. *J. Educ. Psychol.*, *100*(4), 798.

Javed, H., Jeon, M., & Park, C. H. (2018). Adaptive framework for emotional engagement in child-robot interactions for autism interventions. In *2018 15th International Conference on Ubiquitous Robots (UR)* (pp. 396–400). IEEE.

Javed, H., Lee, W., & Park, C. H. (2020). Toward an automated measure of social engagement for children with autism spectrum disorder–a personalized computational modeling approach. *Front. Robotics AI*, *7*, 43.

Jeong, S., Breazeal, C., Logan, D., & Weinstock, P. (2018). Huggable: The impact of embodiment on promoting socio-emotional interactions for Young pediatric inpatients. In *CHI '18, Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3173574.3174069.

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Hum.-Comput. Interact.*, *19*(1–2), 61–84.

Kanda, T., Sato, R., Saiwaki, N., & Ishiguro, H. (2007). A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on Robotics*, *23*(5), 962–971. http://dx.doi.org/10.1109/TRO.2007.904904.

Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., & Göksun, T. (2018). Social robots for early language learning: Current evidence and future directions. *Child Dev. Perspect.*, *12*(3), 146–151. http://dx.doi.org/10.1111/cdep.12277.

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 *HRI '15, Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 67–74). New York, NY, USA: ACM, IEEE.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., et al. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. In *ACM/IEEE International Conference on Human-Robot Interaction, Vol. Part F1271* (pp. 82–90). http://dx.doi.org/10.1145/2909824.3020229.

Komatsubara, T., Shiomi, M., Kanda, T., Ishiguro, H., & Hagita, N. (2014). Can a social robot help children's understanding of science in classrooms? In *Proceedings of the Second International Conference on Human-Agent Interaction* (pp. 83–90). ACM.

Konijn, E. A., & Hoorn, J. F. (2020). Robot tutor and pupils' educational ability: Teaching the times tables. *Comput. Educ.*, *157*, Article 103970. http://dx.doi.org/10.1016/j.compedu.2020.103970, URL https://www.sciencedirect.com/science/article/pii/S0360131520301688.

Konijn, E. A., Jansen, B., Mondaca Bustos, V., Hobbelink, V. L., & Preciado vanegas, D. (2021). Social robots for (second) language learning in (migrant) primary school children. *Int. J. Soc. Robotics*, 1–17.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.*, *15*(2), 155–163. http://dx.doi.org/10.1016/j.jcm.2016.02.012.

Kory-Westlund, J. M., & Breazeal, C. (2015). The interplay of robot language level with children's language learning during storytelling. In *HRI'15 Extended Abstracts*: *vol. 02-05-Marc*, *ACM/IEEE International Conference on Human-Robot Interaction* (pp. 65–66). New York, NY, USA: ACM, http://dx.doi.org/10.1145/2701973.2701989, URL http://doi.acm.org/10.1145/2701973.2701989.

Kory-Westlund, J. M., Dickens, L., Jeong, S., Harris, P., deSteno, D., & Breazeal, C. (2015). A comparison of children learning new words from robots, tablets, & people. In *Proceedings of the 1st International Conference on Social Robots in Therapy and Education* (pp. 28–29).

Kory-Westlund, J. M., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., et al. (2017). Flat vs. expressive storytelling: young children's learning and retention of a social robot's narrative. *Front. Hum. Neurosci.*, *11*, 295. http://dx.doi.org/10.3389/fnhum.2017.00295, URL https://www.frontiersin.org/article/10.3389/fnhum.2017.00295.

Laevers, F. (2003). Experiential education: Making care and education more effective through well-being and involvement. In *Involvement of Children and Teacher Style. Insights from An International Study on Experiential Education* (pp. 13–24).

Laevers, F. (2005). *Well-being and Involvement in care settings. A process-oriented Self-evaluation Instrument (SIC's)* (pp. 1–20). Leuven: Kind En Gezin; Research Centre for Experiential Education.

Laevers, F. (2015). *Making care and education more effective through wellbeing and involvement. An introduction to Experiential Education*. Leuven, Belgium: Research Centre for Experiential Education; University of Leuven, Citeseer.

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2014). Empathic robots for long-term interaction. *Int. J. Soc. Robotics*, *6*(3), 329–341.

Leite, I., Henriques, R., Martinho, C., & Paiva, A. (2013). Sensors in the wild: Exploring electrodermal activity in child-robot interaction. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 41–48). IEEE.

Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: A survey. *Int. J. Soc. Robotics*, *5*(2), 291–308.

Ligthart, M., Neerincx, M. A., & Hindriks, K. V. (2019). Getting acquainted for a long-term child-robot interaction. In *Proceedings of the International Conference on Social Robotics* (pp. 423–433). Cham: Springer International Publishing.

Ligthart, M., Neerincx, M. A., & Hindriks, K. V. (2020). Design patterns for an interactive storytelling robot to support children's engagement and agency. In *HRI '20*, *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 409–418). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3319502.3374826.

Linnenbrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs instudent engagement and learning intheclassroom. *Read. Writ. Q.*, *19*(2), 119–137. http://dx.doi.org/10.1080/10573560308223, http://arxiv.org/abs/DOI: 10.1080/10573560308223[arXiv:DOI: 10.1080/10573560308223].

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Hum. Brain Mapp.*, *32*(6), 982–998.

Mubin, O., Bartneck, C., Feijs, L., Hooft van Huysduynen, H., Hu, J., & Muelver, J. (2012). Improving speech recognition with the robot interaction language. *Disruptive Sci. Technol.*, *1*(2), 79–88.

Mulder, H., Hoofs, H., Verhagen, J., van der Veen, I., & Leseman, P. P. M. (2014). Psychometric properties and convergent and predictive validity of an executive function test battery for two-year-olds. *Front. Psychol.*, *5*, 733.

Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., et al. (2020). Engagement in human-agent interaction: An overview. *Front. Robotics AI*, *7*, 92. http://dx.doi.org/10.3389/frobt.2020.00092, URL https://www.frontiersin.org/article/10.3389/frobt.2020.00092.

Perugia, G., Díaz-Boladeras, M., Català-Mallofré, A., Barakova, E. I., & Rauterberg, M. (2020). ENGAGE-DEM: A model of engagement of people with dementia. *IEEE Transactions on Affective Computing*, 1. http://dx.doi.org/10.1109/TAFFC.2020.2980275.

Piaget, J. (1976). Piaget's theory. In *Piaget and His School* (pp. 11–23). Springer.

Rich, C., Ponsleur, B., Holroyd, A., & Sidner, C. L. (2010). Recognizing engagement in human-robot interaction. In *HRI '10*, *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 375–382). IEEE Press.

Roth, W.-M. (2001). Gestures: Their role in teaching and learning. *Rev. Educ. Res.*, *71*(3), 365–392. http://dx.doi.org/10.3102/00346543071003365, http://arxiv.org/abs/DOI: 10.3102/00346543071003365[arXiv:DOI: 10.3102/00346543071003365].

Rudovic, O., Lee, J., Dai, M., Schuller, B., & Picard, R. W. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, *3*(19), 1–11.

Salter, T., Dautenhahn, K., & Bockhorst, R. (2004). Robots moving out of the laboratory-detecting interaction levels and human contact in noisy school environments. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)* (pp. 563–568). IEEE.

Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., & Paiva, A. (2011). Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 305–312).

Serholt, S., & Barendregt, W. (2016). Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction. In *NordiCHI '16*, *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (p. 64). New York, NY, USA: ACM, http://dx.doi.org/10.1145/2971485.2971536.

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, *166*(1), 140–164. http://dx.doi.org/10.1016/j.artint.2005.03.005, URL http://www.sciencedirect.com/science/article/pii/S0004370205000512.

Storli, R., & Sandseter, E. B. H. (2019). Children's play, well-being and involvement: how children play indoors and outdoors in norwegian early childhood education and care institutions. *Int. J. Play*, *8*(1), 65–78. http://dx.doi.org/10.1080/21594937.2019.1580338, http://arxiv.org/abs/DOI: 10.1080/21594937.2019.1580338[arXiv:DOI: 10.1080/21594937.2019.1580338].

Szafir, D., & Mutlu, B. (2012). Pay attention! designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 11–20).

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, *104*(46), 17954–17958.

Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., et al. (2012). Children with autism social engagement in interaction with nao, an imitative robot: A series of single case experiments. *Interact. Stud.*, *13*(3), 315–347.

Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., et al. (2020). Social engagement of children with autism during interaction with a robot. arXiv:2002.12360.

van Minkelen, P., Gruson, C., van Hees, P., Willems, M., de Wit, J., Aarts, R., et al. (2020). Using self-determination theory in social robots to increase motivation in L2word learning. In *ACM/IEEE International Conference on Human-Robot Interaction*.

Vázquez, M., Steinfeld, A., Hudson, S. E., & Forlizzi, J. (2014). Spatial and other social engagement cues in a child-robot interaction: Effects of a sidekick. In M. Heerink, & M. de Jong (Eds.), *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, Vol. 1* (pp. 391–398). Almere: Windesheim Flevoland.

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., et al. (2019). Second language tutoring using social robots: A large-scale study. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Vol. 2019-March* (pp. 497–505). Ieee, http://dx.doi.org/10.1109/HRI.2019.8673077.

Xu, T., Zhang, H., & Yu, C. (2016). See you see me: the role of eye contact in multimodal human-robot interaction. *ACM Trans. Interact. Intell. Syst. (TiiS)*, *6*(1), 1–22.

Zaga, C., Lohse, M., Truong, K. P., & Evers, v. (2015). The effect of a robot's social character on children's task engagement: Peer versus tutor. In A. Tapus, E. André, J.-C. Martin, F. Ferland, & M. Ammi (Eds.), *Social Robotics* (pp. 704–713). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-25554-5_70.

Zaga, C., Truong, K. P., Lohse, M., & Evers, v. (2014). Exploring child-robot engagement in a collaborative task. In *Child-Robot Interaction Workshop: Social Bonding, Learning and Ethics* (p. 3). Instituto de Engenharia de Sistemas e Computadores, Investigação e desenvolvimento em Lisboa (INESC-ID).