Staying engaged in child-robot interaction

de Haas, Mirjam

DOI:
10.26116/fr4m-gq74

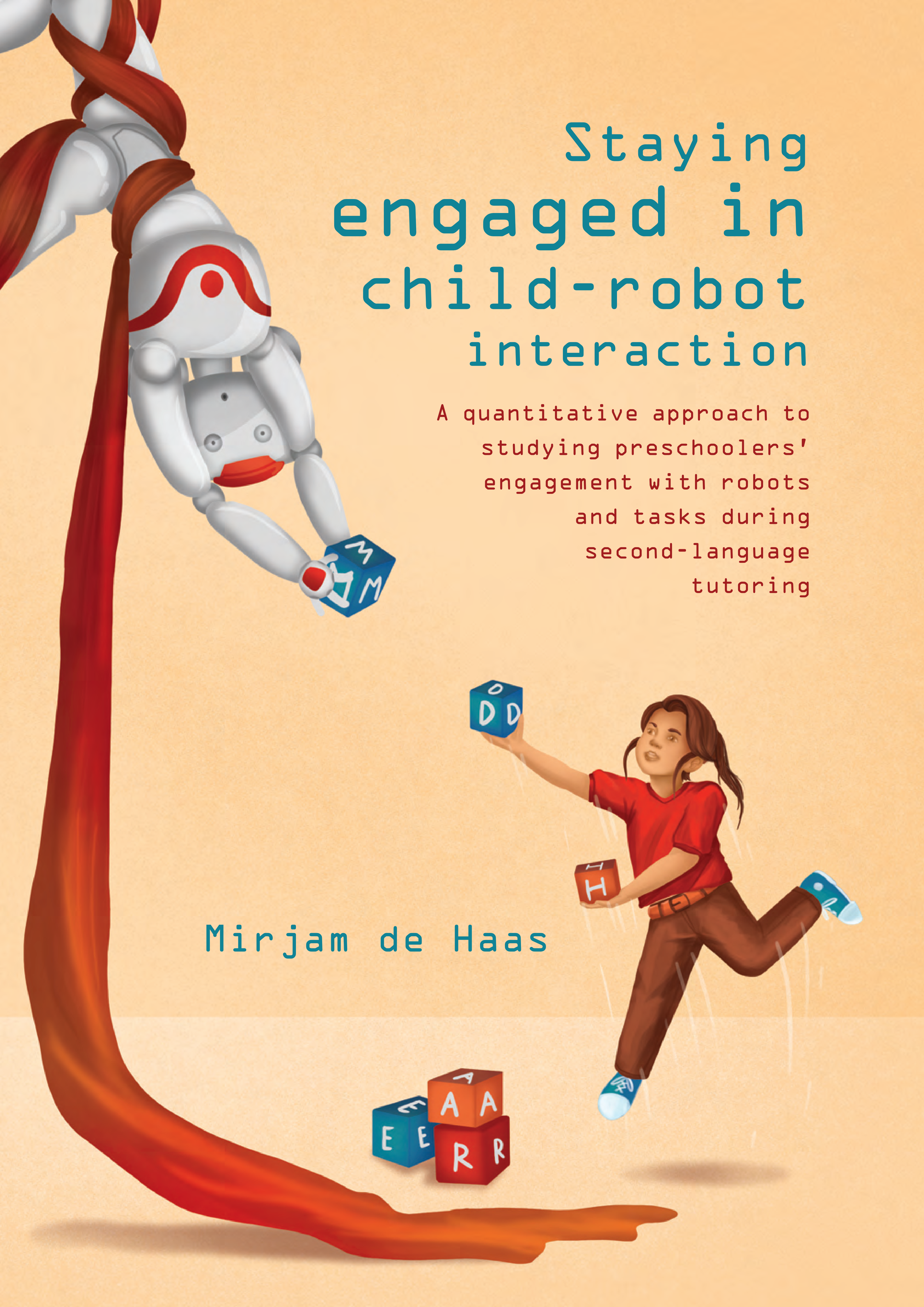
Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Staying engaged in child-robot interaction

A quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring

Mirjam de Haas



Staying engaged in child-robot interaction

A quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring.

Mirjam de Haas

Staying engaged in child-robot interaction: A quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring.

Mirjam de Haas
PhD Thesis
Tilburg University, 2022

Cover design: ©evelienjagtman.com
Printing: Ridderprint | www.ridderprint.nl
ISBN: 978-94-6458-162-1

This dissertation was accomplished with financial support from the European Union Horizon 2020 program (grant number 688014)

SIKS Dissertation Series No. 2022-11
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research school for Information and Knowledge Systems

Staying engaged in child-robot interaction

A quantitative approach to studying preschoolers' engagement with robots
and tasks during second-language tutoring.

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan Tilburg University
op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie
in de Aula van de Universiteit op woensdag 20 april 2022 om 14.00 uur

door

Mirjam de Haas
geboren op 15 maart 1991 te Zoetermeer

Promotor
Prof. dr. Emiel Krahmer, Tilburg University

Copromotor
Dr. Paul Vogt, Hanzehogeschool Groningen

Leden promotiecommissie
Dr. Emilia Barakova, Eindhoven University of Technology
Prof. dr. Tibor Bosse, Radboud University Nijmegen
Prof. dr. Koen Hindriks, Vrije Universiteit Amsterdam
Prof. dr. ir. Pieter Spronck, Tilburg University

Contents

1

Introduction

"Teacher, can I play with the robot now?" This is only one of the questions a child might ask in the future. When I imagine my ideal classroom of the future, each teacher has at least one robot that can be used as a teaching aid in order to support children. When a teacher notices that a child faces difficulties, for example with learning new words from a second language, the teacher may suggest the child to interact with the robot for a while and participate in a one-on-one personalized session and practice the material to be learnt. After this one-on-one session, the child returns to the class and joins the other children for classroom instruction. However, we are only on the doorstep of this future because, before a robot can be used in classrooms, many substantial obstacles need to be overcome in the design of these robot interactions in order to have successful child-robot interactions (importantly, student teachers indicate that they would like to use robots for support, see the text box below. This dissertation, written in the context of the European Horizon 2020 project L2TOR, investigated several design choices and their effects on these child-robot interactions.

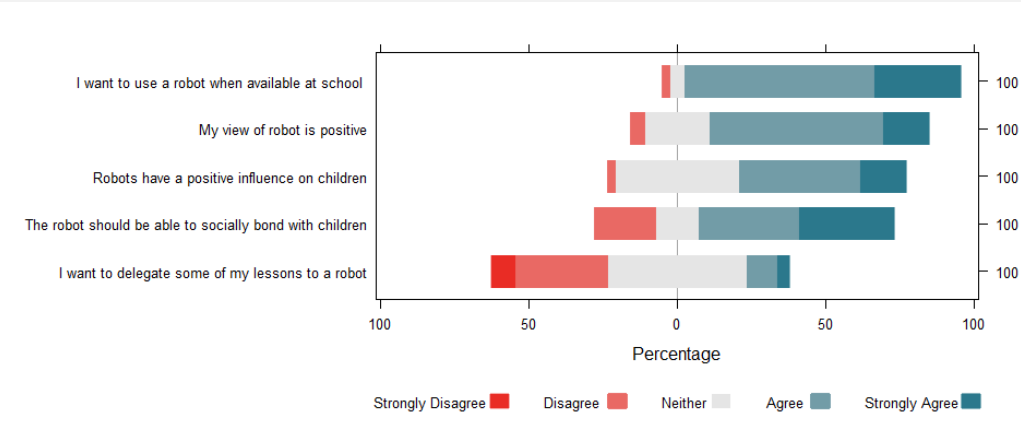
In recent years, the interest in social robots in the field of education has increased substantially (Belpaeme, Kennedy, et al., 2018; van den Berghe et al., 2019). Robots have been used as tutors that support children with various tasks, such as learning to write by handwriting (Hood et al., 2015; Jacq et al., 2016), learning about biodiversity (Ferreira et al., 2017), mathematics and science (Hindriks & Liebens, 2019; Kennedy et al., 2015; Konijn & Hoorn, 2020; Reidsma et al., 2016), or learning a second language (this dissertation, Kennedy et al., 2016; Kory-Westlund & Breazeal, 2015). There are several potential advantages of using a robot in education. First, it may reduce work pressure in the educational field; classes are getting fuller and teachers are not able anymore to provide each child with individual attention (Blatchford & Russell, 2020). Imagine, for example, a child who is struggling with a certain topic; a robot can support this child by practicing smaller tasks related to this topic that can help the child understand it. The teacher can select children who need this extra support and ask them to play one-on-one with the robot before getting back to the group lesson. The robot can provide personalized lessons to support children who face difficulties understanding the material, but also for children who need some extra challenges and are ahead of the other children. Thus, using a robot enables a teacher to spend more time on the other children. This will allow for more inclusiveness in the classroom and enabling children with other difficulties, such as visually impaired children who now have a human assistant providing support, to receive additional support by a robot (Neto et al., 2021). In addition, personalizing the interaction may increase children's learning gain compared to an interaction without personalization (Leyzberg et al., 2014).

Second, robots can provide support in children's native language. This is especially relevant for immigrant children who are learning a new language. While teachers might have some proficiency in some languages, they are not able to provide support in every immigrant child's first language. Since in principle a robot will be able to speak any language, it can provide native language translations to support children in learning a new language. Previous research has shown that Turkish immigrant children in the Netherlands preferred such a bilingual robot to a monolingual one, probably because these children could relate to this bilingual robot and, therefore, felt more connected to this robot (Leeuwestein et al., 2021). This connection might have an effect on their learning because it is possible that children who perceive the robot more as a peer, and therefore benefit from the positive effects of social learning, learn more than when they do not feel this relatedness (Kory-Westlund & Breazeal, 2019a).

Third, the physical presence of the robot can support children's learning and engagement, e.g., by accompanying speech with physical gestures. Moreover, its social nature may stimulate engagement in a variety of ways, including motivating feedback and praising children where appropriate. Gestures can depict meaning of the different concepts and support creating connections between the first and second language. They can also provide extra encouragement and might keep children engaged with the task and with the robot. Keeping children engaged throughout their learning task is important. Children who are more engaged and more motivated to continue with tasks and are more driven to complete the task or to solve a problem (Morgan et al., 1990). In addition, children who are more motivated will learn more during the task because they can remain focused at the task longer than children who are not motivated (Blumenfeld et al., 2006; Halliday et al., 2018). To keep children engaged, using appropriate feedback is important. Feedback can correct a learner but also motivate the learner. Using praise to show children they are performing well can be used to keep up their engagement and can help with children's learning (Cameron & Pierce, 1994).

A challenge in the area of human-robot interaction (HRI) is that robots are a novelty for most people and especially children. This often results in children being highly motivated at the start of a robot interaction and very engaged with the robot and task. However, after some time, the novelty of interacting with the robot starts to wear off and children may become less interested, especially if the robot repeats the same task (Kanda et al., 2004; Leite et al., 2014). It is, therefore, important that more HRI-research focuses on long-term studies, after the novelty effect is gone. In this dissertation, therefore, we address this by looking at the effect of the robot's behavior on the children's engagement during multiple sessions.

Before robots can be used in classrooms, the opinion of teachers is very valuable. It is important to investigate whether teachers see an added value in using robots for teaching. To explore the opinion of student teachers, we asked 27 student teachers to answers questions in a survey about the use of robots at schools. We first introduced the student teachers to the Softbank Robotics NAO robot (see Figure 1.1) and demonstrated some of the robot's abilities, such as several dances and some poses. Moreover, we explained that even though this particular robot has some limitations, such as not being able to move its fingers separately from each other, there are other robots that can and that research is rapidly moving forward. We then asked the student teachers questions about the role of the robot and how they would use the robot in the classroom if they had a robot available. Their answers showed that student teachers are mainly positive about the use of robots in education and they are eager to use a robot. They mostly agree that a robot has a positive influence on children. However, they are not sure whether they want to delegate lessons to a robot.



Moreover, we questioned the student teachers what subjects would be suitable for the robot to teach and their answers were mainly focused on the sciences like physics and mathematics, or languages, such as Dutch and English. The creative subjects such as Art and Physical education were considered less favorable. It is interesting that physical education was less in favor because robots have been used in the past in elderly care and rehabilitation clinics to practice movements with patients (e.g., Assad-Uz-Zaman et al., 2019). However, this can presumably be attributed to the characteristics of the NAO robot being small, slow, and not as capable in executing sports or holding art supplies such as brushes, which is more important with children than with the elderly.

The largest obstacle student teachers reported was the robot's lack of responsiveness when dealing with children (85%), and they indicated that this should be solved for robots to be used at schools. Finally, the participants answered that they do not expect that the robot will be used in schools in the near future as the majority expected that it would take 5-10 years or even longer before social robots are used in schools. Our survey is not the only time that the opinion of teachers is asked (see, for example, Smakman et al., 2021), this survey was one of studies in which teachers respond positively toward the use of social robots while also stating some concerns that need to be addressed.

1.1 Social robots in education

The past decade has seen a rapid development of using social robots for educational purposes, with studies often highlighting the potential of robots as educational tools over more traditional means (reviews by Belpaeme, Kennedy, et al., 2018; van den Berghe et al., 2019; Mubin et al., 2013). One review on robots in education concluded that robots can be effective in teaching children certain skills, but mainly for specific tailor-made lessons (Belpaeme, Kennedy, et al., 2018). When comparing robots to humans, robots seem to have a similar effect as human teaching on children's learning gains, but the effect on children's affective outcomes, such as perception of the robot and engagement, is often larger than on learning gains (Belpaeme, Kennedy, et al., 2018). It is moreover important to note that these studies were tailor-made, meaning that they were specifically designed to test different robot behaviors and do not yet qualify as lessons that can be deployed in actual education. Furthermore, these studies often rely on small sample sizes and measure learning gains over one single robot session, which may not be enough to compensate for the initial engagement learners get when seeing the robot for the first time (van den Berghe et al., 2019; Leite, Martinho, & Paiva, 2013). Before robots can be used in classrooms, there is a need to test robots over multiple sessions and with larger sample sizes.

One of the advantages of a robot compared to a whiteboard or tablet is that children can interact with robots socially in a physical modality, which is noted to be an important factor in second language (L2) learning (Mubin et al., 2013). The physical presence of the robot can be helpful for children in understanding physical topics, more than when they would be interacting with a virtual character. Davison et al. (2021) took advantage of this fact and used a robot to explain different science-related tasks to children, such as explaining how gravity works. The robot's task was to provide the instructions and feedback and the children were asked to answer science-related questions on a tablet placed in front of them. Their findings showed that children successfully interacted with the robot and that they had a more positive mindset with the robot than without the robot, when only learning with a computer. This form of embodiment provides extra support during learning in that the robot can use its arms and body to point at and move toward objects in the physical world and, thus, allowing children to learn more. Another study by de Wit et al. (2018) used iconic gestures, depicting animal shapes, to teach 5-year-old children animal names in a second language. The results showed that children interacting with a robot using iconic gestures remembered more words over time than children interacting with a robot using no gestures. Comparable results were found by Alemi et al. (2015), who investigated 12-year-old children with the robot acting as an assistant teacher using iconic gestures during the explanation of the L2 concepts. The students learned the concepts more effectively, and in addition, showed less anxiety toward using L2. Thus, it seems that the robot's embodiment has a positive effect on children's learning gains.

Moreover, in contrast with more traditional educational tools, social robots can use their humanoid appearance to act out behaviors similar to those of human teachers, while simultaneously helping to keep up the children's motivation (Chang et al., 2010). Appropriate robot behavior for educational purposes, however, may be difficult to design and implement; not only due to technical limitations but also because the intuitive communication with a child that comes naturally to teachers, such as using the appropriate type of feedback for different types of users and situations, may be difficult to realize in a robot tutoring system. Therefore, various studies have investigated different behaviors for robots and their effect on language learning (Gordon et al., 2016; Saerbeck et al., 2010; Kory-Westlund & Breazeal, 2015). For instance, Gordon et al. (2016) compared four-to-six-year-old children learning new words with a robot tutor that personalized affection toward the children with one that did not. The children learned the new vocabulary faster and showed higher valence for the personalized robot. In a similar vein, Kennedy et al. (2016) compared a robot that was verbally more expressive with a non-social robot in a child-robot interaction. The children performed better on the post-test than on the pre-test in both conditions. Therefore, learning was not affected by the different social behaviors, which the authors attributed to the social robot being *too* expressive and hence more distracting for their child learners. It is therefore important to investigate which robot behaviors can keep children engaged with the task and which can contribute to children's learning gain.

Furthermore, this humanoid appearance allows the robot to be perceived as a peer or friend rather than a machine. Children are more inclined to treat the robot as a buddy and anthropomorphize the robot (Lemaignan, Fink, & Dillenbourg, 2014). Presenting the robot as a peer can have an effect on how children treat the robot and will likely create a social bond between child and robot (van Straten, Peter, & Kühne, 2020) which can affect the interaction and children's learning. This social bond with a peer robot has been linked to contribute to children's learning outcomes (Kory-Westlund & Breazeal, 2019b) and is therefore important to take into account while designing robot tutor interactions. The extra advantage of presenting the robot as a peer is that breakdowns might be more acceptable (Vogt et al., 2017).

Taken together, these earlier studies suggest that the behavior of a robot should be carefully designed. Of course, the robot should not be distracting for children, but instead it should be able to build a social bond with the children to create common ground, and to foster engagement. In this dissertation, we study how to achieve this.

1.2 Engagement

Engagement is an important aspect of educational human-robot interactions and can provide information about the state of the learner. There are many definitions of engagement (Oertel et al., 2020) and the most commonly used definition in the HRI field is one by Sidner et al. (2004) who defines engagement as "the process by which two (or more) participants establish, maintain and end their perceived connection" (page 141). The concept thus describes the interaction between two individuals and the interplay between them. If one of them becomes less interested, the process will be considered less successful. Therefore, engagement can be seen as a social process, in the context of this thesis between a child and a robot. It is often the case in HRI that researchers use the term engagement for social engagement between robot and child (*robot engagement*[1]). However, a (child) learner can also be engaged with the task itself. When a child-robot interaction involves a task, it can be expected that the child's attention will be shared between the robot and with the particular task. At this point, a different type of engagement occurs, not a social engagement with just the robot, but rather an engagement in which the task is central: *task engagement*. Hence, in the case of robot interactions in which an extraneous device such as a tablet is used, it is important to also measure task engagement (Zaga et al., 2014). However, even in the case when the task involves

[1] We prefer to use the term 'robot engagement' instead of 'social 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.

different objects, such as blocks or a book, children's attention might be toward those objects and is it important to measure separate robot engagement and task engagement.

Robot engagement and task engagement are likely to be related because the robot plays a large role in the task by acting as the tutor. However, the two engagement types can influence children's learning differently. Robot engagement might, for example, distract the child from the task and therefore reduce their learning gain (Kennedy et al., 2016). In contrast, task engagement is important for learning because the more engaged children are with a task, the more frequently they will perform the task and the more they may learn. It is, therefore, important to look at the relation between both engagement types and children's second-language learning gains.

Engagement may influence not only *short-term* learning but also *long-term* learning. It can have a positive effect on short-term learning because it can result in the child paying more active attention to the task and, therefore, the child may have better concentration during the task (Morgan et al., 1990) which can result into better learning gain during the task. In the long run, engagement can result in children coming back to the lesson more regularly because they enjoyed it (Dörnyei, 1998), but it might also result in them recalling the lesson better and processing the learning materials more thoroughly. This in turn can lead to better retention of the task.

In earlier work, it was generally found that children were relatively highly engaged when interacting with a robot, especially at the beginning of the interaction (Leite, Martinho, & Paiva, 2013). This is often attributed to the aforementioned novelty effect (Kanda et al., 2004), whereby children experience an early peak of engagement due to the novelty and the attractiveness of the robot. After the children get accustomed to the robot, this initial peak will disappear and consequently the robot becomes less interesting. Therefore, it is important to create robot behaviors that not only stimulate children's interest and prolong their engagement but also to test these robot behaviors beyond the effect of the initial novelty with studies involving multiple sessions, as we will do in this dissertation.

Engagement can be measured in multiple ways. Often researchers use methods such as interviewing the participants after the experiment with questionnaires, using behavioral data from videos or physiological measurements such as EEG (e.g., Perugia et al., 2020; Alimardani & Hiraki, 2020). All these methods have advantages and disadvantages, especially when studying children. Self-report measurements are not suitable for younger children, physiological measurements can be too invasive for children. Therefore, studies including children often make use of video analyses. These video data can be processed in multiple ways (either using manual annotations, or automatic ones, often relying on machine learning or deep learning methods). Manual annotations are time consuming, but have the advantage that they are transparent and can focus on those aspects of child behavior which is deemed to be most important for engagement by, for example, teachers. Machine learning and deep learning techniques require a lot of data, something that is most often not available. As a result, researchers sometimes focus on one single aspect of engagement, such as eye gaze, that showed to successfully detect adult participants' disengagement during a robot interaction (Ishii et al., 2013; Nakano & Ishii, 2010). When interviewing preschool teachers for engagement indicators, gaze away was largely part of an indication for boredom and inattentiveness (Schodde et al., 2017). This suggests that the role of eye gaze in engagement detection is large, because it can show the learner's focus point. However, the role of eye gaze as an indicator for engagement is still poorly understood.

1.3 Feedback

Feedback is one way to stimulate engagement and to increase children's learning gain. When children learn something, they receive feedback from their teacher. This feedback helps the learner to understand and to correct their mistakes, but also helps the learner to understand that their progress is going in the right direction. It is evident that feedback helps regarding children's L2 acquisition (Mackey & Oliver, 2002). Mackey & Oliver (2002) conducted an experiment in which children interacted three times with an adult who either gave feedback or proceeded with the story when children gave incorrect answers. The children who received feedback progressed much faster than the children without feedback.

Both *positive* feedback and *negative* feedback can influence children's learning. Generally, positive feedback can increase children's motivation and, therefore, encourage their eagerness to continue with a task which leads to increased learning gains (Blumenfeld et al., 2006; Hattie & Timperley, 2007). Moreover, it goes without saying that negative feedback can also impact children's learning. For example, children who are corrected may try to improve themselves the following instance. Negative feedback and positive feedback can also be combined, for instance when the child will first answer incorrectly, will receive a correction, will correct their first answer, and then receive positive feedback to show that the second answer was correct. This can result in a better learning gain for the child, but possibly also a higher engagement because the child might feel more confident during the task, which can increase the attention to the task and the pleasure during the task, and therefore, results in a higher engagement level. However, it is essential that feedback is perceived as sincere and that it is not provided in every situation, which may cause it to seem less sincere.

It is not only the feedback type but also the person providing the feedback that can influence children's engagement and learning gains. During day-to-day life, children are provided with feedback from everyone around them: from their parents, from their peers, from teachers, other family members, etc. These groups tend to provide feedback in different manners. Teachers typically use generic positive feedback to engage children (Brophy, 1981). Their feedback is usually more focused on the children themselves than in their behavior or performance mainly because teachers know the children's personalities, and thus, they will respond when the children do well. Peer feedback, by contrast, can be more direct than teachers' feedback (Mackey & Oliver, 2002).

All children respond differently to feedback provided by different social groups. For example, one study showed that, in the first grade, boys tend to perform better after peer feedback and girls tend to perform better after adult feedback (Henry et al., 1979). The opposite pattern was seen in fifth-grade children (Dweck & Bush, 1976); fifth-grade girls tend to perform better after peer feedback and fifth-grade boys after adult feedback. Considering that children respond differently to feedback provided by these two social groups, it is likely their reaction will also change for feedback by other social groups, such as feedback provided by the robot. Therefore, it is important to consider the feedback style and the way it is perceived by the learner. In general, the way in which robots provide feedback can be based on how human teachers provide feedback. In this dissertation, we will study which forms of human-inspired robot feedback are most beneficial for children's engagement and learning gains.

1.4 Research context and scope

The work presented in this dissertation was conducted as part of the L2TOR project (www.l2tor.eu), which was funded by the European Union's Horizon 2020 program. The L2TOR consortium was made up of six scientific institutions (Tilburg University, Utrecht University, Bielefeld University, Ghent University, Plymouth University and Koc University) and two industrial companies (Aldebaran Robotics (later Softbank Robotics) and Zora robotics). Each partner took responsibility for one aspect in the development of the robot-assisted tutoring interaction. Tilburg University was primarily tasked with the link between the educational lessons and the technical implementation of the interactions and effect of the robot's behavior on these interactions. The research reported in this dissertation is, there-

Figure 1.1: The NAO robot used in all studies reported in this dissertation

fore, related to this task and focused on how particular design choices of the robot affected the children's engagement.

The aim of the L2TOR project was to develop a long-term interaction for teaching a second language to preschool children with the help of a humanoid robot tutor. This project focused on preschool children, because learning a second language at an early age is important for later academic success (Hoff, 2013; Leseman et al., 2019; Woumans et al., 2016). However, it also provides more challenges in terms of experimenting. Young children have, for example, a shorter attention span than adults (Betts et al., 2006).

The robot used in the L2TOR project is the Softbank Robotics NAO robot (see Figure 1.1 for a picture of the robot). This robot was used in all studies reported in this dissertation. The main feature of this robot is its abstract human-like appearance. It has 25 degrees of freedom, which allowed us to design different types of gestures. The robot is often used in these types of research (Belpaeme, Kennedy, et al., 2018) and there is much support for it. However, during this project the robot's speech recognition was technically deemed not to be ready yet to support speech recognition for preschool children in order to have autonomous interactions (Kennedy et al., 2017). Therefore, for some of the studies in this dissertation, a tablet was used as a mediator between child and robot and most studies were semi autonomous; meaning that there was an experimenter present who had to continue the experiment after voice tasks via a computer interface.

1.4.1 Open science

One of the hallmarks of the L2TOR project was its emphasis on Open Science (github.com/l2tor). The majority of the experiments conducted in Tilburg are uploaded to this repository including the engagement coding scheme used in this dissertation. Several studies were also preregistered via AsPredicted, and datafiles can be found via Dataverse. Preregistration and open access of anonymized data and programming code has become more common practice over the years. However, our studies were among the first in HRI that started to use these practices. The project resulted in many publications over the year, that are all openly accessible via the L2TOR website. These publications vary from speech recognition (or the lack of speech recognition) for young children (Kennedy et al., 2017), the effect of adaptivity and gestures on children's learning (de Wit et al., 2018) to individual differences between children (van den Berghe et al., 2021).

1.5 Structure of the dissertation

This dissertation bundles four studies that explore the effect of feedback and gestures on engagement and on second-language tutoring by a robot over multiple sessions. We investigated the effect of the interaction features on two types of engagement: engagement with the learning task (*task engagement*) and engagement with the robot (*robot engagement*), and on anthropomorphism. Each chapter reports on an individual study that has been published or that is under review as a full paper in an international peer-reviewed journal. Although the chapters are connected to one another, they all contain a separate abstract, introduction, methodology, and discussion. As a result, some overlap between texts may occur.

1.5.1 Insights in measuring engagement

Although the concept engagement, both task and robot, is broadly studied within HRI, there is no consensus of the definition nor of the way of measuring it (Oertel et al., 2020). HRI researchers agree, however, that it is a complex concept and includes many single components. One of these components is eye gaze, especially because children's gaze can show where children's attention is directed to and can easily be automatically measured with video recordings. Therefore, the first research question we will address is:

> **Research question 1**: Can children's eye gaze be used to monitor their task engagement and robot engagement?

In order to answer this question, we first explored the relation between children's eye gaze and task engagement and robot engagement. In Chapter 2, we investigated whether preschool children's eye-gaze direction can predict children's task engagement and robot engagement. Moreover, in Chapter 3, we continue to explore the difference between task engagement and robot engagement and describe our coding scheme for both engagement types in more detail.

1.5.2 Factors that influence engagement

There are several factors that can influence children's task engagement and robot engagement. In this dissertation, we focused on two important aspects: the use of feedback and the use of iconic gestures. We chose these two aspects because of the substantial role of feedback and gestures in learning a second language. Feedback provides the learner with the correct form and motivation to continue, which can have an influence on engagement. Robotic gestures are shown to support second-language learning in human studies, and might also increase children's engagement.

> **Research question 2:** Do robotic feedback and iconic gestures influence children's task engagement and robot engagement?

In Chapter 2 and Chapter 3, we describe two experiments in which we used different forms of robotic feedback to teach children some vocabulary in a second language. We measured children learning gain and their engagement and compared whether there were differences over time and between conditions. In Chapter 2, we specifically look at the difference between feedback in different robotic roles: peer-like, adult-like, and no feedback with 3-year-old children. In Chapter 3, we first asked student teachers to provide a preferred feedback method and used their recommendations to test teacher-recommended feedback and compared it to feedback they did not recommend and no feedback with 5-year-old children.

1.5.3 Novelty effect

Previous studies suggest that children's engagement with robots decreases over time, the so-called novelty effect (Ahmad et al., 2019; Kanda et al., 2007; Leite, Martinho, & Paiva, 2013; Oertel et al., 2020). This novelty effect describes the effect that new technology first increases children's engagement while it drops again later because children become less interested. To keep children engaged and motivated over time, it is important to discover what type of robot behavior can keep children engaged over time.

> **Research question 3:** How do children's task engagement and robot engagement develop over time?

In Chapter 3, we describe an experiment that measured children's task engagement and robot engagement over three sessions and investigated the influence of different kinds of robotic feedback on children's engagement over time. Chapter 4 describes the influence of robot gestures on children's task engagement and robot engagement over seven sessions.

1.5.4 Relation between children's engagement and word learning

From human-human studies, we know that children's engagement is related to their learning gain. However, in the child-robot studies, this effect has not been extensively studied yet. Therefore, our fourth research question is:

> **Research question 4:** What is the relation between children's task engagement and robot engagement and their second-language learning gain?

In Chapters 2, 3 and 4, we investigated the relation between children's word knowledge and their task engagement and robot engagement. Chapter 2 describes an experiment with toddlers, Chapter 3 an experiment where the robot providing different types of feedback to preschoolers, and Chapter 4 an experiment in which the robot used different types of gestures when teaching the second language to preschoolers.

1.5.5 Children's perception of the robot

One aspect that can influence children's engagement is the children's social bond with the robot. Research showed that this social bond is stronger when children anthropomorphize the robot (Kory-Westlund & Breazeal, 2019a) which can result in children perceiving the robot more as their peer. This peer relation can increase children's learning gain because peer learning has shown to be beneficial for learning gain (see for a review Topping, 2005), either via direct support from the peer, or via increasing the motivation and confidence of the learner and therefore result in higher learning gain.

> **Research question 5:** How do children's perceptions of the robot develop over time when interacting with a robot tutor and is it related to their L2 learning gain?

Chapter 5 describes the degree to which children attributed human-like properties to the robot and whether those were related to the effectiveness of the word learning training.

1.5.6 Discussion

Finally, this dissertation provides an answer to the research questions and a general discussion on the key findings of all the studies described and will discuss limitations of our studies and recommendations for future work.

2

On the interplay between eye gaze, robotic feedback and engagement in a short-term child-robot interaction

Based on: de Haas, M., Vogt, P., & Krahmer, E. (2021). When Preschoolers Interact with an Educational Robot: Does Robot Feedback Influence E ngagement? *Multimodal Technologies and Interaction, 5*(12), 77. https://doi.org/10.3390/mti5120077

Abstract In this chapter, we examine to what degree children of 3-4 years old engage with a task and with a social robot during a second-language tutoring session. We specifically investigated whether children's task engagement and robot engagement were influenced by three different feedback types by the robot: adult-like feedback, peer-like feedback and no feedback. Additionally, we investigated the relation between children's eye-gaze fixations and their task engagement and robot engagement. Fifty-eight Dutch children participated in an English counting task with a social robot and physical blocks. We found that, overall, children in the three conditions showed similar task engagement and robot engagement; however, within each condition, they showed large individual differences. Additionally, regression analyses revealed that there is a relation between children's eye-gaze direction and engagement. Our findings showed that although eye gaze plays a significant role in measuring engagement and can be used to model children's task engagement and robot engagement, it does not account for the full concept and engagement still comprises more than just eye gaze.

2.1 Introduction

In recent years, the interest in using robots for educational purposes has increased substantially (Belpaeme, Vogt, et al., 2018; van den Berghe et al., 2019) due to the growing numbers of students in classrooms, shrinking school budgets and the fact that robots can possibly exhibit social behaviors that can benefit children's learning (Belpaeme, Vogt, et al., 2018). One application in the educational domain that utilizes robots is second-language (L2) learning (van den Berghe et al., 2019; Kanero et al., 2018) in which robots are often used as tutors to support children's L2 acquisition. In order to be successful as a robot tutor, the robot should be able to engage the children in order to motivate them during the task. The aim of this chapter is to investigate children's engagement during a second-language tutoring session with a social robot.

Engagement plays an important role in learning. Engaged children are more motivated and are more likely to continue longer with their learning tasks than disengaged children. The more time children are actively interacting with a certain task, the more children can learn from that task. The engagement of primary and middle school children has frequently been studied, being linked numerous times to children's academic performances (e.g., Roorda et al., 2011; Fredricks et al., 2004).

In human-robot interaction (HRI), it is common that people to initially be highly engaged but quickly start to become less engaged as the task continues due to its repetitive and the novelty of the task wearing off. This novelty effect is observed in both the engagement with the robot as a social partner (robot engagement) and in the engagement with the task within the robot interaction (task engagement) (Oertel et al., 2020). This distinction between task engagement and robot engagement is important because children can be engaged with the learning task in front of them, but not with the tutor, or vice versa. Both engagement types can have an influence on children's learning (Oertel et al., 2020), although the results are inconsistent (e.g., Kennedy et al., 2015). Previous studies on HRI typically only measured the engagement with the robot and not with the task given to the participant (Ahmad et al., 2019; de Wit et al., 2018; Oertel et al., 2020). The reason for this is that researchers are specifically interested in the effect of their manipulation, which is often a result of the robot's behavior. However, it is also worth examining task engagement (Zaga et al., 2015), because this may reveal whether the learner's engagement decreased in response to the experimental task or the robot's behavior.

There are several methods that are able to stimulate and maintain children's task engagement and robot engagement, and one of them is feedback. Providing children with the correct form of feedback is essential, as different children seem to respond better to different feedback types (Byrne, 1987). On the one hand, positive feedback can motivate children, keep them engaged during a task and can activate their learning behavior (Masgoret & Gardner, 2003). On the other hand, for other children, it might decrease their performance, when the children receive this feedback too often, it becomes too repetitive and, as a result, the children become less engaged (Flink et al., 1992). Children can also respond differently to negative feedback, especially young children. As young children (preschoolers) quickly absorb all the information around them and rely on their environment for (correct) input, they tend to benefit more when they receive corrective feedback than adults would (Mackey & Oliver, 2002). In contrast, younger children might be more sensitive to explicit negative feedback than older children, particularly when it guides them to notice errors (Lyster & Saito, 2010; Mackey et al., 2003; Oliver, 2000). Moreover, negative feedback can lead to frustration which can decrease children's motivation to fully participate in the task and therefore decrease children's task engagement (D'Mello & Graesser, 2012).

This chapter aims to investigate children's task engagement and robot engagement during a second-language (L2) learning task with a robot by specifically focusing on the role of the robot's feedback on the children's engagement. Moreover, we investigate the role of children's eye gaze on their task engagement and robot engagement. In the following sections, we provide an overview of earlier work on engagement in child-robot interaction and feedback in education. We then explain the design of the experimental study and, finally, we will present the results and discuss our findings.

2.2 Background

2.2.1 Engagement

Numerous studies across the HRI field focus on engagement. After all, the key to continuing to use robots in different fields is when people remain interested in robots, especially over time. For many people, robots are something new and hence interesting. However, over time, this interest may change. Consequently, engagement is widely studied and frequently, when researchers refer to the concept of engagement, a variety of definitions are used. The most commonly used definition in HRI is by Sidner et al. (2005), who defined engagement as "the process by which individuals in an interaction start, maintain and end their perceived connection to one another" (page 141), but there are others who argue that it is more than a cognitive process and explain engagement as a multidimensional concept of a cognitive dimension (such as attention), an affective dimension (such as emotions) and a behavioral dimension (such as the execution of tasks) (Trowler, 2010; Zaga et al., 2014). Although there has been a large variation in the definition of engagement and in how it has been studied, there is an agreement that engagement is a multidimensional concept.

Children are normally very engaged with the robot, but this quickly decreases over time which has been shown in numerous experiments (e.g., Tanaka & Matsuzoe, 2012; Kanda et al., 2004; Leite et al., 2014). It is, therefore, important to understand which robot behavior can lead to a positive effect on children's engagement. Many studies have investigated the effect of robot behavior on children's engagement, looking at different robot behaviors such as the robot's gestures (de Wit et al., 2018), expressiveness of the voice (Kory-Westlund, Dickens, et al., 2017), or the role of the robot (Zaga et al., 2015; Chen et al., 2020). de Wit et al. (2018) found a positive effect of robot gestures on 5-year-old children's engagement. Kory-Westlund, Dickens, et al. (2017) found that 5-year-old children were more engaged with a robot exhibiting expressive behaviors. A recent study showed that 5- to-7-year old children who interacted with a robot acting as a peer showed more affect during the interaction than when interacting with a robot acting like a tutor (Chen et al., 2020).

A disadvantage of these studies is that they focus on engagement with the robot instead of the task. However, it is possible that task engagement is more important for learning. A follow up study by de Wit et al. (2020) found a positive effect of robot gestures on children's robot engagement but not on task engagement nor learning gain. Moreover, Zaga et al. (2015) investigated task engagement during a robot tutoring session. In their experiment, they compared a robot behaving as a peer and a tutor and found that children were more engaged in the task and solved the task faster with the peer-like robot than with the tutor-like robot.

Similarly to how there are different definitions of engagement, there are also different methods for measuring engagement (Oertel et al., 2020). For adults, questionnaires can be used as self-reported measures. This can be useful to determine participants' own reflection of the interaction. Unfortunately, questionnaires only provide a total rating *after* the interaction and not *during* the interaction, and are difficult to use with children. Other methods are based on video or audio data and measure participants' output behaviors, such eye gaze, head movements (nodding), verbal utterances and facial expressions or a combination of these behaviors (Inoue et al., 2018; Rich et al., 2010; Ishii et al., 2013). Eye gaze is especially important, because it can indicate where the participant's attention is directed and can relatively easily be measured automatically, making it ideal for real-time engagement tutoring interactions.

Some studies have examined the role of eye gaze within engagement (Ishii et al., 2013; Nakano & Ishii, 2010; Rich et al., 2010). Nakano & Ishii (2010) and Ishii et al. (2013) used automatic gaze direction to initiate probing questions by the robot whenever the participant looked away from the robot, indicating disengagement. This showed to have a positive effect on the participants' non-verbal and verbal behaviors. However, they concentrated their study on the social interaction between the robot and participant and did not investigate what happens when a task is in front of the participant. This can result in different eye-gaze behaviors such as looking away from the robot more often. Rich et al. (2010) combined mutual gaze and joint attention to determine the participant's engagement and this combination increased the participant's attention to the robot. However, these studies do not differentiate between robot engagement and task engagement and it is possible that they actually measured participants' engagement with the robot. Moreover, these studies did not investigate whether it is feasible to monitor eye gaze with children and whether children's eye gaze relates to engagement. Although eye gaze only focuses on one aspect of engagement, it undoubtedly plays a role because it can show the direction of the participant's attention which is one of the three dimensions of engagement according to (Trowler, 2010). It does not, however, explain the whole concept. Therefore, it would be interesting to examine how large the role of this single element is and whether this role is large enough to successfully predict children's task engagement and robot engagement during a L2 learning tutoring session.

2.2.2 Feedback

Research on L2 learning has demonstrated the importance of feedback and engagement on children's language learning performance on human-human studies (Mackey & Silver, 2005). While the role of feedback has extensively been studied on human-human interaction, in the field of child-robot interaction it is largely understudied (see Ahmad et al., 2019; Hindriks & Liebens, 2019). In order to design social robots as effective L2 tutors, it is therefore important to investigate how a social robot should provide feedback to optimize children's engagement.

In general, educational robots are designed based on how human teachers interact with their pupils; however, in classroom settings, children not only receive feedback from their teachers but also from their peers. Teachers normally provide a combination of positive and negative feedback. They use explicit positive feedback to encourage the children and they recast the children's answers to provide corrections as a type of implicit negative feedback (Lyster & Ranta, 1997). Their positive feedback can result in children becoming more engaged with the task and when they are fully engaged, they learn faster and continue longer with the task (Oxford & Shearin, 1994; Dörnyei, 1998; Kanda et al., 2004). The use of recasts during L2 learning provides a subtle way to correct the children's mistakes. In the case of a recast, the adult will repeat the utterance, but rephrase the incorrect part into a correct one (e.g., when a child had said: "The cat is jumping", he/she may be corrected by the adult's utterance: "Ah right, the *dog* is jumping"). The use of recasts is additionally intended to avoid providing demotivating comments found in explicitly negative feedback.

Children do not only receive feedback from their teachers, they also receive feedback from peers in their classroom (Mashburn et al., 2009). In contrast to the implicit feedback that adults provide, children tend to use more explicit language ("No, you are wrong!") (Mackey & Oliver, 2002). It has been argued that explicit feedback can have a more substantial impact on learning than implicit feedback (Okita & Schwartz, 2013). However, the potential side effect of providing explicit feedback is that children's engagement decreases. As shown, all of these different forms of feedback provide children with the correct information but in a different manner, and consequently, these different forms may have a different influence on the children's engagement. In addition, children might have feedback preferences, where one child might remain more engaged with explicit feedback, while implicit feedback might stimulate engagement more for another child.

Given that learners do not exclusively receive feedback from adults, the design of robot feedback on the basis of the teacher's feedback might not always be the most optimal for children's development. For example, research has shown that the presence of a peer improves learning potential (Mashburn et al., 2009), and as a result, some researchers have argued that educational robots might work better when presented as peers, especially since children may treat the robot as a peer rather than a teacher in long-term interactions (Kanda et al., 2004). Therefore, it might be better to design feedback provided by a robot based on children's peer interactions.

The use of feedback in child-robot interaction studies has not been extensively studied. Adult-robot interaction studies showed that participants listened more to negative feedback provided by a robot than negative feedback provided by computers (Midden & Ham, 2014), that participants learned more words during an L2 learning task from a robot providing only negative feedback than a robot providing only positive or no feedback (de Haas & Conijn, 2020), and positive feedback positively influenced adults' acceptance of the robot as instructor (E. Park et al., 2011), and increased adults' motivation (Midden & Ham, 2014). However, it is difficult to relate these results to children because children learn differently to adults.

In child-robot interactions, most studies report the use of praise or various types of negative feedback, such as introducing a doubt ("Are you sure?") instead of negative feedback (Mazzoni & Benvenuti, 2015), providing hints ("I think it was the other one") (Gordon et al., 2016) or providing children with an extra attempt after an incorrect answer (de Haas et al., 2016), but these studies did not investigate the effect of these feedback utterances on children's learning gain or engagement. Hindriks & Liebens (2019) investigated a robot providing feedback to 7-9-year-old children solving mathematical problems. The feedback was based on the most likely error children made solving the problem (e.g., forgetting to add one number). Their results showed that these extra explanations were appreciated by children who found these mathematical problems difficult, but children who solved problems quickly became impatient by the robot's feedback. These findings indicate that feedback technique preferences might different for certain groups. Only one study has investigated the use of robot feedback in L2 learning (Ahmad et al., 2019). In this study, children of eight to ten years old played a game with a robot on a tablet. The robot provided either positive emotional feedback, negative emotional feedback or neutral feedback. Ahmad and colleagues found that the robot providing positive emotional feedback positively influenced children's learning gain and their social engagement with the robot. These studies did not investigate, however, children's engagement with the task and the question arises as to whether the same effects can be found with younger children, who are shown to rely more on the experimenter than the robot (Baxter et al., 2017). Moreover, the study of Ahmad used a tablet as an interaction medium between the robot and child. However, a disadvantage of using a tablet is that it can play a large role in the interaction (Vogt et al., 2019; Konijn et al., 2021) and reduce the children's attention to the robot tutor, which can lead to a decrease in children's robot engagement and their learning gain. It is, therefore, interesting to investigate the influence of robot feedback on children's task engagement and robot engagement during a robot tutoring session without a tablet present.

2.2.3 This study

In this study, children received a tutoring session from a social robot and learned how to count in English using physical blocks. We investigated whether children were more task-engaged and more robot-engaged with a robot providing adult-like feedback (implicit negative feedback and explicit positive feedback), peer-like feedback (explicit negative feedback) or no feedback, and whether eye-gaze direction can predict task engagement or robot engagement. Finally, we investigated the relation between children's task engagement and robot engagement with children's learning gain. We addressed the following hypotheses:

H1 a) Children are more *task-engaged* with a robot that provides feedback than with a robot that does not provide feedback.
We expect that children's task engagement will be higher when children receive feedback because the feedback will make them aware of their mistakes. This awareness can lead to a more successful completion of the task and children's success will result in a higher task engagement.

b) Children are more *robot-engaged* when the robot provides adult-like feedback than in the other two conditions. We expect this result because the adult-like feedback is the only condition that provides positive feedback, which is shown to increase children's motivation and can increase children's robot engagement (Kluger & deNisi, 1996; Hattie & Gan, 2011). We expect that this effect will mainly contribute to children's robot engagement because the robot is providing the positive feedback and children might like the robot more due to these positive expressions.

H2 a) Eye gaze toward the blocks and the robot has a positive relation with children's *task engagement* and children's eye gaze elsewhere has a negative relation with children's *task engagement*.
We expect that this is because the task involves both the robot as an instructor and the blocks because the children have to manipulate these blocks during the task.

b) Children's eye gaze toward the robot will have a positive relation with *robot engagement* and the other eye-gaze directions will have a negative relation with *robot engagement*.
We expect that only eye gaze toward the robot will have a positive relation with robot engagement, because when you communicate and, therefore, engage with a robot as a social partner, this is often accompanied by mutual eye gaze with this social partner (Mwangi et al., 2018) and other studies that detected disengagement with the robot (Nakano & Ishii, 2010; Ishii et al., 2013; Rich et al., 2010) when participants looked away.

2.3 Method

A between subject design with three conditions was employed for this study. Children received either adult-like feedback, peer-like feedback or no feedback. The robot behavior remained the same through the conditions except for the robot's feedback.

2.3.1 Participants

A total of 58 native Dutch children (M_{age} = 3 years and 6 months, SD = 4 months) participated in this study. All children attended a preschool or childcare in the Netherlands. For all children, the parents signed an informed consent form to give permission. The participants were randomly distributed over the three conditions. Four children indicated that they wanted to stop participating during the experiment and therefore stopped the experiment prematurely and were removed from the data. This resulted into the following distribution:

1. Adult-like feedback ($N = 21$, M_{age} = 3 years and 6 months, 12 boys and 9 girls);
2. Peer-like feedback ($N = 18$, M_{age} = 3 years and 6 months, 10 boys and 8 girls);
3. No feedback ($N = 19$, M_{age} = 3 years and 7 months, 13 boys and 6 girls)

Exact age data for four children are missing and are not included in the age calculation. The study was conducted in accordance with the Declaration of Helsinki, and received ethical approval from the Research Ethics committee of Tilburg School of Humanities and Digital Sciences.

2.3.2 Robot tutoring session

The interaction was completely in Dutch, except for the target words, which were in English (the target words are italicized in this section to indicate which words were spoken in English). The aim of the session was to teach children to count from one to four in English. Before the tutoring session, children participated in a group introduction and received a pre-test. The tutoring session started with the robot teaching the children the four counting words using different training tasks. These training tasks varied from repeating the target words, counting various body parts of the robot to building a tower with blocks and counting the height of the tower. For instance, the robot would ask the child to build a tower and to count together how tall the tower is: "Shall we count together in English how tall this tower is? Repeat after me: *one, two, three, four*." (in Dutch: "Zullen we samen tellen hoe hoog de toren is in het Engels? Zeg mij maar na: *one, two, three, four*."). All target words were repeated three times during these training tasks. After this concept binding of the target words, the robot and child went of the different target words with the use of the four blocks. For each target word, the robot asked the child to collect a certain number of blocks using an English counting word: "I'm going to say in English how many blocks you should grab: *three*" (in Dutch:"Ik ga in het Engels zeggen hoeveel blokken jij mag pakken: *three*"). The order of the target words was fixed and was, therefore, the same for each child. Each target word was asked only once during these practice rounds to reduce the duration of the experiment. Once the child collected the blocks, the robot provided feedback (only in the adult-like and peer-like feedback conditions) and continued with the next instruction. After all words were practiced, the robot and child concluded the session with a Dutch children's dance.

2.3.3 Experimental conditions

The children received either adult-like feedback, peer-like feedback, or no feedback, see for an example Table 2.1:

1. In the adult-like feedback condition, the robot used explicit positive feedback for correct answers and implicit negative feedback for incorrect answers. A correct answer would invoke a facial expression using colored eye-LEDs and positive verbal feedback ("That is right, *three* means three in English"). For an incorrect answer, corrective feedback was provided ("*three* means three"). After receiving negative feedback, children could try again ("You should take *three* blocks"), after which the robot would again provide feedback. This negative feedback was, at most, provided twice for every target word, which means that during the experiment, every child was able to receive negative feedback eight times and positive feedback four times. In case the child gave more than two incorrect answers, the robot still provided positive feedback and continued to the next instruction. For both positive feedback and negative feedback, the robot repeated the English target word, which increased children's exposure to the target words.

2. In the peer-like feedback condition, the robot did not provide positive feedback but only provided explicit negative feedback. This explicit negative feedback was based on children's feedback during peer interaction (Long, 2006). Similar to the adult-like feedback condition, children could try again twice after receiving negative feedback. After a correct answer, the robot would continue to the next step without any feedback.

3. In the no feedback condition, the robot did not provide any feedback and just continued the game with the blocks after children collected the correct or incorrect number of blocks.

Table 2.1: An example of the robot's feedback in the different feedback conditions.

Condition	Correct answer	
	Dutch	English
Adult-like	Dat is goed! Three betekent drie in het Engels.	That is right! *Three* means three in English
Peer-like	-	-
No feedback	-	-
	Incorrect answer	
	Dutch	English
Adult-like	Three betekent drie, je moet drie blokken pakken. Probeer opnieuw	*Three* means three, you should take three blocks. Try again
Peer-like	Dat is fout! Je moet drie blokken pakken. Probeer opnieuw.	That is wrong! You should take three blocks. Try again.
No feedback	-	-

2.3.4 Materials

Experimental setting

The experiment took place in multiple preschools and childcare centers in the Netherlands. At each location, the experiment room was a classroom that the children were familiar with, but not in use by the school. The Softbank Robotics NAO robot was used, which is commonly deployed in experiments with children. Moreover, four blue blocks were used. We chose to use blocks in our experiment because preschool children are used to playing with

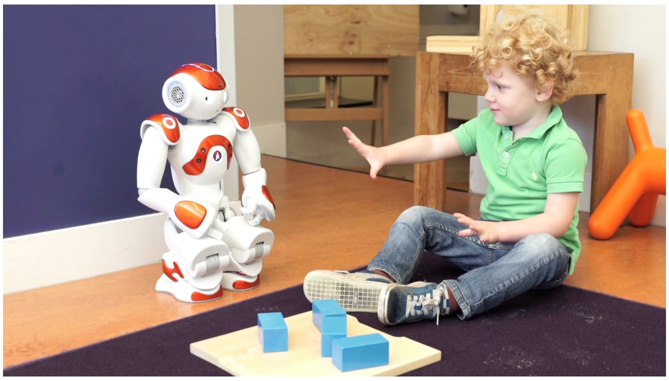


Figure 2.1: The setup of the experiment.

blocks, and children learn how to manipulate and handle blocks to enhance their visual-spatial skills (Casey et al., 2008). The children sat on the ground with the crouched robot, approximately 40 cm from each other, with the blocks in between. (see Figure 2.1 for the experimental setting). The children were positioned so they could not see the corridor and, therefore, could not see other children passing by the room. The children were filmed from two viewpoints: one camera was positioned in front of the child to record his or her face and one camera was sideways to record the social interaction between robot and child. Two experimenters were present during the interaction to operate the robot and to provide reassurance for the children if necessary. While the experimenters sometimes instructed the child to perform a task if required, they were careful not to provide feedback.

Pre-test

Before the child started the tutoring sessions, his or her Dutch and English knowledge of the four target words was tested. The experiment leader asked the child to collect a number of blocks and repeated this for every target word (e.g., "Can you give me four blocks?"). This process was first completed in Dutch to test their L1 knowledge of the target words, and then in English to test their L2 knowledge. The same blocks were used as during the tutoring session; however in this case six blocks were used instead of four to reduce the chance of guessing. For both Dutch and English, the experimenter noted how many target words the child already knew in both languages. The experimenter did not provide feedback between the words, and only continued with the next target word.

Post-test

The post-test was the same as the pre-test, however it was only conducted for the English target words. The experimenter used six blocks and asked the children to collect the number of blocks that was equal to each of the four target words.

2.3.5 Procedure

Group introduction

The study consisted of two group introductions and one tutoring session. One week before and in the morning of the experimental day, group introductions were given to familiarize the children with the robot and build up trust and rapport with the robot (Vogt et al., 2017). All children in the classroom participated during the first group introduction, but only the children that participated in our experiment joined the second introduction. Both introductions were the same, and during these we highlighted some of the similarities of the robot with people to establish common ground, since this can have a positive effect on the learning outcome (Kanda et al., 2004). For example, we explained that the robot has arms and legs just like people have and can express emotions through its eye-LEDs. The robot and children would then dance a familiar Dutch children's song. We never forced the children to participate; if they declined they could sit in a quiet corner and watch from a distance.

Experiment

After the child was brought to the experiment room, the experimenter tested the child's prior knowledge of the target words in both Dutch and English, as described in Section 2.3.4. The pre-test was carried out in the same room as the tutoring session, but at some distance from the robot. After the pre-test, the child was asked to sit in front to the robot. During the experiment, the two experimenters remained in the room at a distance to discourage children from looking at them. When children looked at the experimenters or asked them for help, the experimenters redirected the children's attention back to the robot. When a child displayed signs of discomfort, the experimenters comforted the child and tried to make him or her more relaxed. For some children, the experimenters remained close to the children and helped them during the beginning of the interaction. In the case of four children, the experiment was stopped and these children were brought back to their classrooms.

After the robot tutoring session was completed, the experiment finished with an English post-test. When this post-test was completed, a short debriefing was conducted. During this debriefing, the experimenters repeated all of the target words and their translation to ensure that children had learned the correct translation. Finally, the child was brought back to their classroom. The duration of the experiment was approximately 10 to 15 minutes.

2.3.6 Engagement and gaze coding

We manually coded three different aspects of the interaction: task engagement, robot engagement and children's eye-gaze direction. Not the whole interaction was coded; instead we chose two video fragments: one two-minute fragment at the beginning of the interaction, and a two-minute fragment at the end of the interaction. The gaze coding was only completed for the two-minute fragment at the end of the interaction due to time constraints. The two video fragments of the interaction were chosen to code different aspects in the interaction, with the first fragment being the moment when the robot started to teach the children English, and the other fragment being a moment in the end of the interaction when the robot and child started to play with the blocks. These fragments resulted in 116 video fragments for 58 children.

Engagement coding

Both task engagement as well as robot engagement were rated on a Likert scale from one to five, including half points, with one as a low level of engagement and five being highly engaged. We based our engagement coding scheme on an existing coding scheme named ZIKO [1] (Laevers, 2005). This coding scheme is used in children's day cares to measure, among other things, children's engagement to improve the day care activities. We adapted the scheme to include specific cues for our own experiment, such as attention toward the experiment leader instead of the robot and blocks. Children were fully task-engaged, when they were completely "absorbed" in the robot-block activity, when they showed to be open for new information, were very motivated and listened to the tasks. Robot engagement described children's engagement with the robot as a social partner and focused more on the interaction itself than on the task. Each engagement level had specific cues for the rater.

A *high task engagement* had cues to look for such as: looking at the task and robot, actively answering and grabbing blocks, listening for new instructions and being fully committed to the task. In contrast, a *low task engagement* was indicated by fiddling, not performing, and playing with objects not related to the task (e.g., their shoes). A *neutral task engagement* was determined as the child executing the tasks but not being fully immersed in them.

A *high robot engagement* had cues such as: looking at the robot, having an open body posture toward the robot, having spontaneous conversations with the robot. A *low robot engagement* had cues such as: turning away from the robot. A *neutral robot engagement* had cues such as touching the robot without meaning. For all specific cues and information, see the coding scheme on Github [2].

Ten percent of the data were coded by two raters and their inter-rater agreement was considered moderate using the intraclass correlation coefficient (ICC_{task} = .75, 95% CI [.05, 93], ICC_{robot} = .64, 95% CI [.16, .88]) (Koo & Li, 2016).

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

[2] https://www.github.com/l2tor/codingscheme.

Eye-gaze coding

We coded children's eye gaze toward different directions in order to measure their visual attention using ELAN (Wittenburg et al., 2006). We analyzed the same fragments as engagement, but only the second fragment when children also used the blocks for the interaction. In particular, we coded children's eye gaze in five different directions: the robot, blocks, experimenter, elsewhere and unknown. The latter direction unknown (0.71%) was not included in the analysis. Eye gazes that were shorter than one second were excluded and added to the nearest annotation, as a short glance would not change the children's focus point. For the analyses, we calculated the duration for each category. To assess inter-rater reliability for this categorical data, 10% of the videos were coded by a second annotator, yielding a Fleiss' Kappa of .83 which is considered a very good agreement.

2.3.7 Analyses

We investigated children's task engagement and robot engagement over the session and the conditions. We measured the two engagement types in the beginning of the session and the end of the session.

To inspect the normality of the engagement data, Q-Q plots were plotted and the Shapiro-Wilk test was conducted. Both the plots and Shapiro-Wilk tests showed a non-normal distribution of the task engagement and robot engagement. Consequently (Field et al., 2012), we conducted two robust two-way mixed design ANOVAs with 20% trimmed means and the feedback condition as a between-subject variable and the test moment (beginning and end of session) as a within-subject variable on both engagement types. We used the "WRS2" R package to conduct this analysis (Mair & Wilcox, 2020).

To investigate the relation between children's eye-gaze direction and their engagement, multiple regression analyses of task engagement and robot engagement were performed using four predictors: duration of eye-gaze toward the blocks, the robot, the experimenter and elsewhere. The assumptions of non-multicollinearity were checked using variance inflation factor (VIF) statistics (Hutcheson & Sofroniou, 1999). Several models were analyzed, from which the best model was chosen.

We investigated the effect of the different feedback types on children's learning gain. A Q-Q plot and a Shapiro-Wilk tests showed a non-normal distribution of the learning gain. Therefore, we conducted a robust mixed design ANOVA with 20% trimmed means to test the effect of the tutoring session and feedback on children's word knowledge.

Finally, we investigated the relation between engagement and learning using a Pearson correlation analysis.

2.4 Results

First, we will report on the effects of the experimental conditions on children's task engagement and robot engagement. Next, we will discuss the relation between children's eye-gaze direction and their engagement. Finally, we will report the effect of the three feedback conditions on children's learning gain and the relation of learning gain and engagement.

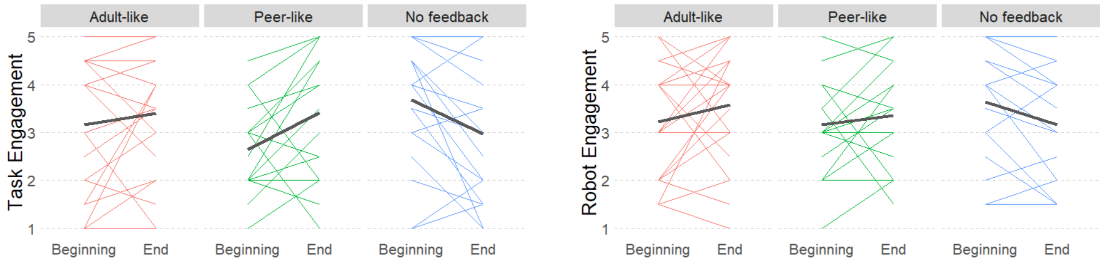
2.4.1 Engagement

To begin, we investigated whether task engagement and robot engagement were related. Task engagement and robot engagement were correlated ($r(218) = .70, p < .001$), indicating that children who scored higher on task engagement also scored higher on robot engagement.

Task engagement

We investigated whether the three experimental feedback conditions had an effect on children's task engagement. We expected that children would be more task-engaged when the robot was providing feedback. Figure 2.2a shows that there were large individual differences in children's task engagement over time, conditions and between the individual children. Some children became more task-engaged over time (48%), other children became less task-engaged over time (38%) and other children were equally engaged in the beginning of the session as in the end (14%). When looking at the graph, on average children scored higher than a neutral task engagement (3.0), except at the beginning of the session for the peer-like feedback condition.

We carried out a robust two-way mixed design ANOVA using trimmed means on children's task engagement with condition as between factor and test moment (beginning of the session and end of the session) as within factor. In contrast to our expectations, there was no significant difference between children in the different conditions ($F(2, 22.67) = 0.58, p = .57$), nor was there a significant difference over time ($F(1, 33.37) = 0.12, p = .73$). However, there was a significant interaction effect between condition and test moment ($F(2, 23.17) = 6.89, p = .004$). This interaction effect is illustrated in Figure 2.2a, children's task engagement in the peer-like and in the adult-like feedback conditions increased during the session and in the no feedback condition their task engagement decreased over time.



a: Task engagement   b: Robot engagement

Figure 2.2: The individual children's (a) task engagement scores and (b) robot engagement scores over the three conditions in the beginning and end of the tutoring session. The dark lines show the averages of the children's engagement scores.

Robot engagement

Similar to children's task engagement, children's robot engagement varied for each condition in the beginning and end of the session (see Figure 2.2b). It decreased for 45% of the children, increased for 43% of the children and remained the same for 12% of the children. To investigate whether their robot engagement was different for the three feedback conditions, we conducted a robust two-way mixed design ANOVA using trimmed means on children's robot engagement with condition as between factor and test moment (beginning of the session and end of the session) as within factor. We expected that children would be more robot-engaged when interacting with a robot providing adult-like feedback than the other two conditions. Contrary to these expectations, there was no significant difference between children in the different conditions ($F(2, 22.57) = 0.16, p = .85$), nor a difference over time ($F(1, 31.57) = 0.01, p = .93$). Similar to task engagement, there was a significant interaction effect between condition and test moment ($F(2, 22.36) = 3.88, p = .04$). This means that children's robot engagement was influenced by the three feedback conditions and the moment in the session. Figure 2.2 shows that children's robot engagement increased during the two feedback conditions and decreased in the no feedback condition. This increase in robot engagement appeared to be less strong than with task engagement.

2.4.2 Duration of eye-gaze directions as engagement predictor

Next, we investigated whether the duration of children's different eye-gaze directions had a relation with children's engagement. Table 2.2 shows the duration in seconds toward the robot, human experimenter, the blocks and elsewhere in the different conditions. Overall, children spent the most time looking at the blocks, followed by the robot, the experimenter and they spent the least time looking elsewhere. To investigate the relation between the duration of each eye-gaze direction and engagement, we carried out a linear regression to predict the role of eye gaze on children's task engagement and robot engagement.

Table 2.2: The mean duration in seconds for the children's eye-gaze direction divided into each feedback condition (SD between brackets).

Condition	Robot	Blocks	Experimenter	Elsewhere
Adult-like	45.1 (21.0)	57.1 (15.3)	13.5 (12.0)	2.7 (3.4)
Peer-like	38.1 (24.2)	57.2 (30.1)	15.4 (15.2)	2.4 (3.3)
No feedback	31.6 (19.3)	56.0 (18.5)	18.6 (17.5)	5.6 (7.9)
Overall	38.1 (21.9)	56.8 (21.6)	15.9 (15.0)	3.6 (5.5)

Task engagement

Table 2.3 shows the different regression analyses we performed. **Model 1** included all eye-gaze directions and when combined, these explained a significant proportion of the variance of task engagement ($F(4, 50) = 16.13, p < .001, R^2_{adj} = .53$). However, when checking for multicollinearity, we found that the duration that children looked toward the blocks and toward the robot were highly related (VIF scores: $blocks = 6.45$, $robot = 6.48$) and strongly correlated ($r = -0.69, p < .001$). Following (Hutcheson & Sofroniou, 1999), we combined these two directions by taking the sum of the two directions and using their total duration (*blocks and robot*) in a new model. **Model 2** also explained a large proportion of variation ($F(3, 51) = 21.54, p < .001, R^2_{adj} = .53$) with acceptable VIF values (VIF scores: $blocks\,and\,robot = 3.86$, $experimenter = 3.22$, $elsewhere = 1.54$). As an alternative to using the total duration in both eye-gaze directions, **Model 3**, we removed the predictor with the highest VIF value (Hutcheson & Sofroniou, 1999), which was the duration children looked at the robot. In this alternative model, the duration that children looked at the blocks did not contribute significantly to the prediction. Hence, we removed this variable from the model. The resulting **Model 4** significantly explained 48% of the task engagement's variance and did not perform better than the other models. Therefore, the best model was Model 2

($R^2_{adj} = .53$) and the resulting regression equation was:

$$Eng_{task} = 8.89 - 0.04 \times Gaze_{blocks\ and\ robot} - 0.09 \times Gaze_{experimenter} - 0.10 \times Gaze_{elsewhere} \quad (2.1)$$

where Eng_{task} is task engagement, $Gaze_{blocks\ and\ robot}$ the duration in seconds of eye-gaze toward the blocks and the robot, $Gaze_{experimenter}$ is the duration toward the experimenter and $Gaze_{elsewhere}$ is the duration that children looked elsewhere.

Robot engagement

For robot engagement we used a similar approach as for task engagement. We performed different multiple regression models to predict children's robot engagement using the duration of children's eye gaze toward the blocks, the robot, the experimenter and elsewhere. Similarly to task engagement, the model containing all variables explained a significant proportion of the variance of children's robot engagement (see Table 2.4 for the models). **Model 1** showed that 58% of the variance in children's robot engagement can be explained by the duration in which children looked at the four different eye-gaze directions. However, both the duration that children looked in the direction of the robot and in the direction of the blocks did not significantly contribute to the model and could therefore be removed. We ran three further models: **Model 2** without children's eye-gaze direction toward the robot, **Model 3** without children's eye gaze toward the blocks and **Model 4** without both the eye gaze toward the robot and blocks. Model 4 contained the fewest predictors, but also explained the lowest variance of the four models ($R^2 = .45$). Despite that the other two models (3 and 4) were the same regarding the variance ($R^2 = .58$), we prefer the model containing gaze toward to the robot instead of blocks because this model (**Model 3**) shows the positive relation between eye-gaze direction to the robot and robot engagement.

Table 2.3: Regression analyses summary for the duration (s) that children looked in different directions predicting children's task engagement.

Eye-gaze direction	Coefficient	*SE*	*VIF*	*t*	*P*
Model 1					
constant	8.93	1.66		5.39	$< .001$
robot	-0.04	0.02	6.48	-2.55	.01
blocks	-0.04	0.02	6.45	-2.86	.01
experimenter	-0.09	0.02	3.23	-5.75	$< .001$
elsewhere	-0.10	0.02	1.55	-4.84	$< .001$
Model 2					
constant	8.89	1.65		5.40	$< .001$
blocks and robot	-0.04	0.01	3.86	-2.78	.01
experimenter	-0.09	0.02	3.22	-5.80	$< .001$
elsewhere	-0.10	0.02	1.54	-4.83	$< .001$
Model 3					
constant	4.87	0.48		10.20	$< .001$
blocks	-0.01	0.01	1.14	-1.23	.22
experimenter	-0.06	0.01	1.07	-6.03	$< .001$
elsewhere	-0.07	0.02	1.06	-3.92	$< .001$
Model 4					
constant	4.34	0.21		20.80	$< .001$
experimenter	-0.06	0.01	1.00	-5.88	$< .001$
elsewhere	-0.07	0.02	1.00	-3.71	$< .001$

Model 1: $Eng_{task} = \alpha + \beta \times$robot $+\beta \times$ blocks$+\beta \times$experimenter$+\beta \times$elsewhere
$F(4, 50) = 16.13, p < .001, R^2_{adj} = .53, RSE = .29$

Model 2: $Eng_{task} = \alpha + \beta \times$blocks and robot $+\beta \times$experimenter $+\beta \times$elsewhere
$F(3, 50) = 21.54, p < .001, R^2_{adj} = .53, RSE = .29$

Model 3: $Eng_{task} = \alpha + +\beta \times$blocks $+\beta \times$experimenter $+\beta \times$elsewhere
$F(3, 51) = 17.45, p < .001, R^2_{adj} = .48, RSE = .30$

Model 4: $Eng_{task} = \alpha + \beta \times$experimenter $+\beta \times$elsewhere
$F(2, 52) = 25.18, p < .001, R^2_{adj} = .47, RSE = .30$

Table 2.4: Regression analyses summary for the duration (s) that children looked in different directions predicting children's robot engagement.

Eye-gaze direction	Coefficient	*SE*	*VIF*	*t*	*P*
Model 1					
constant	4.87	1.31		3.71	$< .001$
robot	0.01	0.01	6.48	0.61	.55
blocks	-0.01	0.01	6.45	-1.18	.24
experimenter	-0.05	0.01	3.23	-4.13	$< .001$
elsewhere	-0.04	0.02	1.55	-2.61	.01
Model 2					
constant	5.63	0.36		15.69	$< .001$
blocks	-0.02	0.00	1.14	-4.15	$< .001$
experimenter	-0.06	0.01	1.07	-8.10	$< .001$
elsewhere	-0.05	0.01	1.06	-3.59	$< .001$
Model 3					
constant	3.35	0.28		11.81	$< .001$
robot	0.02	0.01	1.14	3.98	$< .001$
experimenter	-0.04	0.01	1.14	-5.33	$< .001$
elsewhere	-0.03	0.01	1.01	-2.36	$< .001$
Model 4					
constant	4.29	0.18		24.09	$< .001$
experimenter	-0.05	0.01	1.00	-6.32	$< .001$
elsewhere	-0.04	0.02	1.00	-2.34	.02

Model 1: $Eng_{robot} = \alpha + \beta\times$robot $+\beta\times$blocks $+\beta\times$experimenter$+\beta\times$elsewhere
$F(4, 50) = 19.48, p < .001, R^2_{adj} = .58, RSE = .22$

Model 2: $Eng_{robot} = \alpha + \beta\times$blocks $+\beta\times$experimenter$+\beta\times$elsewhere
$F(3, 51) = 26.17, p < .001, R^2_{adj} = .58, RSE = .22$

Model 3: $Eng_{robot} = \alpha + \beta\times$robot $+\beta\times$experimenter$+\beta\times$elsewhere
$F(3, 51) = 25.31, p < .001, R^2_{adj} = .58, RSE = .22$

Model 4: $Eng_{robot} = \alpha + \beta\times$experimenter$+\beta\times$elsewhere
$F(2, 52) = 23.35, p < .001, R^2_{adj} = .45, RSE = .25$

The resulting regression equation for robot engagement was:

$$Eng_{robot} = 3.35 + 0.02 \times Gaze_{robot} - 0.04 \times Gaze_{experimenter} - 0.03 \times Gaze_{elsewhere} \quad (2.2)$$

where Eng_{robot} is robot engagement, $Gaze_{robot}$ the duration in seconds that children looked at the robot, $Gaze_{experimenter}$ is the duration that children looked at the experimenter and $Gaze_{elsewhere}$ is the duration of children's eye gaze elsewhere.

2

2.4.3 Learning gain

Next, we examined whether different forms of feedback influenced children's word knowledge. Table 2.5 reveals that children on average knew between one and two words after the session, but standard deviations are high. Children performed above chance level in the pre-test (chance level = .16, $W = 4879, p < .001$) and post-test (chance level = .16, $W = 4945, p < .001$). A robust mixed-design ANOVA with 20% trimmed means and with children's word knowledge as dependent variable and with condition as between factor and the two test moments (pre- and post-test) as within variable showed that children did not know significantly more words ($M = 1.43, SD = 0.95$) after the session than before the session ($M = 0.98, SD = 0.65; F(1, 17.71) = 3.76, p = .07$). There were no significant differences between the three conditions ($F(2, 14.94) = 1.81, p = .20$) nor a significant interaction effect between conditions and test moment ($F(2, 14.94) = 0.05, p = .95$). This showed that children did not know significantly more target words after the session than before the session, independent of the condition.

Table 2.5: The children's average word knowledge scores on the pre-test and post-test for the three conditions (SD between brackets).

Condition	Pre	Post
Peer-like	1.18 (0.7)	1.61 (0.9)
Adult-like	0.90 (0.4)	1.38 (1.0)
No feedback	0.91 (0.8)	1.33 (0.9)
Total	0.98 (0.7)	1.43 (0.9)

2.4.4 Relation learning gain, task engagement and robot engagement

Finally, to investigate whether there is a relation between L2 word knowledge and children's task engagement and robot engagement, we performed a Pearson correlation analysis. We did not find any significant correlation between children's learning gain and task engagement ($r(109) = 0.12, p = .21$). Likewise, we did not find a significant correlation between robot engagement and learning gain ($r(109) = 0.16, p = .10$), meaning that children's engagement levels did not have a relation with how many words children learn.

2.5 Discussion

In this chapter, we presented a study in which we investigated the role of robot feedback on toddlers' task engagement and robot engagement, their learning gain and the relation between toddlers' eye-gaze direction and engagement. The children were assigned to one of three feedback conditions: a robot providing feedback like an adult would (adult-like feedback), a robot providing feedback like a peer would (peer-like feedback) and a condition where the robot provided no feedback (no feedback). While task engagement and robot engagement are different concepts, they are moderately correlated and show similar trends. Both engagement types decreased when children did not receive any feedback and increased during the session for peer-like feedback and adult-like feedback. Moreover, for both engagement types there were large individual differences between children. Given these similar trends for the two engagement types, we will first discuss the results combined and then discuss the differences in our findings between these two engagement types.

2.5.1 Engagement

We investigated children's task engagement and robot engagement in the beginning and the end of the tutoring session with the robot. Overall, children were engaged with the task and robot, and their engagement remained approximately the same over time. Contrary to our expectations, there was no main effect of feedback on children's *task engagement* (H1a) nor on children's *robot engagement* (H1b). This unexpected result may be explained by the fact that the robot's behavior did not differ sufficiently in the three conditions. The current study only provided a limited number of exposures to the target words. Hence, there might not have been enough feedback moments in order to observe a significant effect across the different conditions. Although this is a limitation of our design, we did not want to increase the duration of the session because children's attention span at this young age is very short (Betts et al., 2006). In future investigations, it might be recommended to use multiple sessions with these young children, to measure an effect on children's task engagement and robot engagement. There are, however, other possible explanations. The result might also be explained by the fact that the children in our current study were very young. Children undergo a major developmental progress at this age and learn how to think more logically when they get older (Piaget, 1976). It is possible that younger children need more, or other types of feedback than older children. Another possible explanation for our findings is that the individual differences between children are larger than the differences between the conditions. As Figure 2.2 shows, there were many individual differences between children, which is in line with other studies (e.g., Leite et al., 2017). It is possible that some children would have been more engaged with a robot providing peer-like feedback and other children with adult-like feedback. Our study did not include enough participants to investigate these individual differences, and future studies with more participants will need to be undertaken.

Furthermore, there was no main effect of time on task engagement and robot engagement, which is, again, surprising because in previous experiments children's task engagement dropped over time within one session (de Wit et al., 2018; van Minkelen et al., 2020). It is possible that this is due to the duration of the session: our session was shorter than those of de Wit et al. (2018) and van Minkelen et al. (2020) due to the shorter attention span of the children, which might explain the difference between the previous studies and the current one. It is also possible that there was too much variation between children and conditions, that nullified the effects over time. Finally, a specific explanation for the lack of results for *task engagement*, is that the beginning of our task itself (counting together with the robot) was very different than the end (playing with the blocks) and that this variation kept children task-engaged. In the two experiments by de Wit et al.(de Wit et al., 2018, 2020), the task remained the same during the full session and it is likely that children's task engagement dropped due to the lack of variation (Ahmad et al., 2017). Our expectation is that this game variation will mostly influence task engagement; however since we did not investigate this, it is possible that it will also influence robot engagement, e.g., because the robot's instructions are more important during one aspect of the task and as a result children look more at the robot which will increase their robot engagement.

While we did not find a main effect of conditions or time on engagement, we did find an interaction effect of condition and time. When inspecting Figure 2.2, we can observe that children's task engagement increased for both feedback conditions and it decreased in the no feedback condition over time. We saw a similar pattern for robot engagement. It is likely that children in the no feedback condition became less task-engaged and robot-engaged during the session because this condition did not include any feedback whether they completed the tasks successfully or unsuccessfully. The absence of positive confirmation when children accomplished the task may have played a role in their task engagement and robot engagement and therefore may have reduced it (Kluger & deNisi, 1996; Hattie & Gan, 2011). In a similar way to how the absence of corrective feedback to help the children in the rest of the session might have reduced their attention for the learning task, it also possibly increased frustration (Deci & Ryan, 1985) which consequently could have led to task disengagement. Thus, feedback seems to have a positive effect on children's engagement over time.

2.5.2 Duration of eye-gaze directions as engagement predictor

We explored the relation between eye gaze and children's task engagement and robot engagement in order to understand whether this important but single aspect of engagement, can successfully predict task engagement and robot engagement.

Our findings showed that children's eye-gaze direction can explain a large proportion of the variance of both children's task engagement and robot engagement. In particular, children's *task engagement* had a negative relation with the duration children looked at the robot and blocks combined, and with the duration children looked at the experimenter and elsewhere. There were multiple models possible for robot engagement: (1) a negative relation with the duration children looked at the *blocks, the experimenter* and *elsewhere* and (2) a positive relation with the duration children looked at the *robot*, a negative relation with the duration children looked at the *experimenter* and *elsewhere*. These results might seem surprising; however, when looking at the regression equations they can be explained. For children's *task engagement*, all gaze directions were taken into account in the equation. Our expectation (H2a) was that the duration that the children looked at the blocks and at the robot would have a positive relation with children's task engagement and the duration that children looked at the experimenter or elsewhere a negative relation. Our regression equation showed that the duration that children looked at the experimenter and elsewhere would lower the rate of task engagement more (with factors of 0.09 and 0.10 respectively) than the duration that children looked at the blocks and robot combined (0.04). The larger role of looking elsewhere (and perhaps looking at the experimenter) supports previous studies that used eye gaze to detect disengagement and as a cue to initiate different robot behaviors that can re-grab the participant's attention (Ishii et al., 2013). It is possible that children's disengagement (attention away from the task and directed at the experimenter + elsewhere) is easier to detect using eye gaze and can be used in future studies to initiate engagement-increasing behaviors in the robot.

Moreover, for children's *robot engagement*, there were two models performing equally well: a model including the duration that children looked at the blocks, experiment and elsewhere and a model including the duration that children looked at the robot, experiment and elsewhere. The model including robot gaze had a positive relation with robot engagement (H2b), and the model including eye gaze toward to blocks had a negative relation with robot engagement. These can both be explained by examining the regression coefficients. When gaze toward the robot is not included, all other eye-gaze directions have a negative effect on robot engagement. However, when robot gaze is included, this eye-gaze direction has a positive relationship with robot engagement. Therefore, even though both models will explain robot engagement equally well, we prefer to use the model containing eye gaze toward the robot because it follows intuitively that eye gaze at the robot predicts robot engagement.

Our results provide further support for the hypothesis that eye gaze is a good predictor for task engagement and robot engagement and that future studies can use eye gaze for automatic systems to detect engagement. These studies might additionally incorporate the robot's on-board camera to measure children's gazes in order to reduce the extra hardware needed. However, this can be complicated because the robot's head often moves. Eye gaze explained a larger proportion of the variance for robot engagement than task engagement, a possible explanation is that robot engagement is a social engagement, and social interaction is often based on eye gaze toward each other (Mwangi et al., 2018). A note of caution is due here since not all of the variance can be explained by eye gaze (task engagement (53%) and for robot engagement (58%)) which indicates that eye gaze does not predict every aspect of children's engagement (both task and robot). For task engagement elements such as speech, emotional expressions, children's fiddling or children's interaction with the blocks should be included and for robot engagement elements such as speech toward the robot, smiles during the conversation and body posture can be considered as predictors for engagement.

2.5.3 Learning gain

Contrary to our expectations, we found that children did not learn during the interaction nor was this dependent on the condition. Children knew, irrespective of condition, no more target words after the experiment than before the experiment. It is likely that the exposure to each target word was not enough, which reduced the training of target words and therefore children's learning gain. Three- and four-year-old children have a limited attention span of 3 to 4 minutes (Gaertner et al., 2008) and although our experiment lasted already much longer, we did not want to exhaust the children by introducing more repetitions. To create a more successful tutoring session, the exposure to target words should be higher. Moreover, it should be noted that the children's general word knowledge after the sessions was low for all three conditions, which is a result more commonly found after robotic tutoring sessions (van den Berghe et al., 2019). Future studies should look at repeating target words over sessions, and perhaps focus on the words children did not know yet in the earlier sessions instead of repeating all words (creating a more personalized interaction).

There were large individual differences between children: some children learned all the target words and some did not learn any words. These individual differences is in line with previous research (van den Berghe et al., 2021), where we specifically investigated the individual differences between preschoolers learning with a robot and found that the robot gestures benefited children's word knowledge in different ways across children. In our current study, it is possible that some children benefited from the adult-like feedback, and others from peer-like feedback.

2.5.4 Individual differences

As already mentioned, there were large individual differences between children. Interestingly, when looking at children's responses in an exploratory manner, there is an overlapping pattern for each condition. For example in the peer-like feedback condition, although the robot instructed the children to collect a certain number of blocks, a third of the children misunderstood the robot and simply repeated the target word (seven children) or repeated the word while also collecting the blocks (five children). This observation may be explained by the fact that the child had to repeat the word to the robot during the word concept binding phase of the interaction and they got used to repeating the L2 word when the robot used this L2 word. A similar variation was observed in the adult-like feedback condition, instead of collecting the blocks after the robot's instructions, three children built a tower, six repeated the robot and the experimenter had to intervene five times.

Moreover, children frequently requested additional support from the experimenter after the robot's instruction. Some children hardly looked at the robot and always looked at the experimenter while grabbing the blocks or needed additional persuasion to show the blocks to the robot. The experimenter intervened approximately four times during the whole tutoring session after the robot's instructions, this varied from repeating the robot's instruction, asking the children to grab the blocks instead of repeating the words, and instructing the children to pay attention to the robot.

Finally, some children started playing with the blocks and completely ignored the robot. For instance, they started to throw the blocks, to play with their shoes, and even started to play with the microphone close to the robot. The experimenters intervened when this happened and tried to redirect the child's attention to the robot, but some children lost their engagement completely. Occasionally, these children regained focus after the next instruction. This is probably due to the low attention span of this age group. This behavior is unfortunately inevitable, as there will always be children who have little attention for the task. Whether this is due to external factors, such as being fatigued or to the task itself is something that researchers should take into account when designing child-robot interactions.

Taken together, the children's responses after the robot instructions varied considerably. Other studies should, therefore, focus on personalizing the interactions for every child, even with preschool children like in our experiment (Gordon et al., 2016; Leyzberg et al., 2014). In our experiment, we did not personalize the interaction in order to maximize the systematic effect of different feedback types on the children's task and robot engagement, although we did not find any differences.

2.6 Conclusion

Given the potential of social robots for tutors with preschool children, it is important to understand how children can be effectively tutored, while still being engaged with the task and robot. In this study, we investigated the effect of the robot's feedback on young children's task engagement and robot engagement in a second-language tutoring session. The robot either provided feedback as an adult, as a peer or no feedback during the tutoring session. Moreover, we explored the relation between eye-gaze direction and robot engagement and task engagement. Our findings showed that there was an interaction effect between children's engagement and the three feedback conditions. Providing feedback (as a peer and adult) increased children's task engagement and robot engagement during the session, while providing no feedback did not increase the task engagement and robot engagement. Finally, our study shows that children's eye-gaze direction is informative for children's task and robot engagement, which can contribute to automatic engagement measuring systems in child-robot tutoring interactions.

3

The effect of robotic feedback on children's engagement and learning gain during multiple interactions

Based on: de Haas, M., Vogt, P., & Krahmer, E. (2020). The Effects of Feedback on Children's Engagement and Learning Outcomes in Robot-Assisted Second Language Learning. *Frontiers in Robotics and AI 7*, 101. https://doi.org/10.3389/frobt.2020.00101

Abstract To investigate how a robot's use of feedback can influence children's engagement and support second-language learning, we conducted an experiment in which 72 children of five years old learned 18 English animal names from a humanoid robot tutor in three different sessions. During each session, children played 24 rounds in an "I spy with my little eye" game with the robot, and in each session the robot provided them with a different type of feedback. These feedback types were based on a questionnaire study that we conducted with student teachers and the outcome of this questionnaire was translated to three within-design conditions: (teacher) preferred feedback, (teacher) dispreferred feedback, and no feedback. During the preferred feedback session, among others, the robot varied his feedback and gave children the opportunity to try again (e.g., "Well done! You clicked on the horse", "Too bad, you pressed the bird. Try again. Please click on the horse"); during the dispreferred feedback the robot did not vary the feedback ("Well done!", "Too bad") and children did not receive an extra attempt to try again; and during no feedback the robot did not comment on the children's performances at all. We measured the children's engagement with the task and with the robot as well as their learning gain, as a function of condition. Results show that children tended to be more engaged with the robot and task when the robot used preferred feedback than in the two other conditions. However, preferred or dispreferred feedback did not have an influence on learning gain. Children learned on average the same number of words in all conditions. These findings are especially interesting for long-term interactions where engagement of children often drops. Moreover, feedback can become more important for learning when children need to rely more on feedback, for example, when words or language constructions are more complex than in our experiment. The experiment's method, measurements and main hypotheses were preregistered.

3.1 Introduction

A recent trend in education is to have social robots take on the role of educational tutors to support, for example, second-language (L2) learning (Belpaeme, Kennedy, et al., 2018; Vogt et al., 2019; Kory-Westlund & Breazeal, 2015). In educational settings, learning a (second) language typically involves social interactions between the child and the teacher. During these interactions, children constantly receive feedback about their performance. It has been shown that human feedback can have a clear impact on children's learning process and outcomes (Hattie & Timperley, 2007; Wojitas, 1998). Feedback is, therefore, an important part of the social interactions that facilitate language learning, which begs the question what the impact of various feedback types is when feedback is provided by a robot rather than a human.

Throughout many years researchers have investigated how (human) feedback can have an influence on L2 learning. Focusing on children learning a second language, research has shown that receiving feedback benefits children's language development more than receiving no feedback (Mackey & Silver, 2005). Moreover, different types of feedback can help children in several ways. You can, for example, use positive feedback to reward and motivate children when they are correct, or use negative feedback to correct children's language and thereby improve children's learning gain (Hattie & Timperley, 2007).

While there have been many studies about robots for educating children, only few of these have investigated the effects that different types of feedback can have on children's engagement and learning performance (Ahmad et al., 2019; Hindriks & Liebens, 2019). Usually, studies design feedback strategies for robot tutors based on results from educational studies involving only humans without investigating the effect that these strategies have on children's engagement and/or performance (e.g., Gordon et al., 2016; Kennedy et al., 2016; Mazzoni & Benvenuti, 2015; Kory-Westlund & Breazeal, 2015). However, it is not evident that the effect of human strategies will be the same when a robot uses them, because a robot has substantial cognitive and physical limitations compared to a human. For example, robots cannot produce the same facial expressions as humans or humans' subtle cues, thus are limited in providing facial cues that humans use to provide non-verbal feedback (Vogt et al., 2017).

One recent study manipulated non-verbal and verbal feedback based on the child's emotional state (Ahmad et al., 2019). Results showed that children's engagement over time remained relatively high and children's word knowledge increased over time with positive or neutral feedback. While their results suggest that robot feedback can have a positive effect on children's engagement and learning gain, they did not compare different variations of positive and negative feedback or compared it against no feedback. The results of Ahmad et al. (2019) are consistent with findings from human studies and demonstrate that feedback does not only enhance children's language performance, but also engages children. Positive feedback engages because it validates children's answers and thus boosts their confidence (Henderlong & Lepper, 2002; Zentall & Morris, 2010). Similarly, negative (corrective) feedback corrects and teaches the child the correct word which could result in a motivated child. However, both positive and negative feedback can also decrease engagement. On the one hand, too many repetitions of positive feedback can become meaningless for a child and can result in less intrinsic motivation (Boyer et al., 2008; Henderlong & Lepper, 2002). On the other hand, negative feedback can decrease the child's confidence and thereby decrease the engagement between the teacher and child (Wojitas, 1998).

Consequently, if used correctly, feedback can result in increased learning gains. Children become more intrinsically motivated by positive feedback, which increases the children's interest and their task engagement and therefore their skills. These increased skills will motivate the children further and engage the children to a greater extent (Blumenfeld et al., 2006).

This chapter describes a study that investigated how preschool children respond to different types of feedback provided by a robot tutor. In the experiment, children interacted with a humanoid robot tutor in three different L2 sessions, and in each session the children received a different type of feedback. These types of feedback were designed based on a survey among student teachers, resulting in a strategy preferred by these student teachers, a strategy dispreferred by them and a strategy using no feedback at all. We analyzed the effect of these different types of feedback on the children's task engagement and learning gain over time.

3.2 Background

3.2.1 Feedback

Numerous studies have shown that feedback facilitates L2 learning (Hattie & Timperley, 2007; Henderlong & Lepper, 2002; Long, 2006; Lyster & Ranta, 1997). It can help to improve pronunciation, word choice and grammar, and makes it easier for children to understand what is correct or incorrect in the foreign language. Feedback is not only used to correct children, but for example also by teachers to contribute positively to children's own feeling of competence and success and therefore encourage children to continue with a task (Blumenfeld et al., 2006; Hattie & Timperley, 2007). The type of feedback provided, however, matters (Shute, 2008). You can, for example, provide explicit negative feedback by indicating that something is wrong with children's answers, but without specifying what was wrong (e.g., 'That's wrong.'). It is also possible to provide corrective feedback by correcting children's answers or hinting toward it (e.g., "You said runned, but it should have been ran" or "it should not have been runned, but...?"). Prompting children with an extra attempt ("Try again.") is an implicit way of saying something was wrong. Hattie & Timperley (2007) propose a combination of these three types as good way of providing feedback. The combination provides children with explicit notions where the mistake was made, what went wrong and makes them to try again. Nevertheless, sometimes separate feedback is also sufficient. For example, using explicit negative feedback (i.e. stating explicitly that something is wrong) seems to be most beneficial for children who are struggling with a task, such as novel learners (Kluger & deNisi, 1996; Shute, 2008).

Teachers, however, mostly provide negative feedback implicitly by using recasts (i.e. a type of feedback in which the teacher repeats the incorrect phrases in a correct form), but they still try to make sure that children reach their goal (Long, 2006; Lyster & Ranta, 1997). Although these recasts have been found to be used more often than the other feedback types, they seem to be less effective in helping the learner to reach their learning goal. Lyster & Ranta (1997) investigated the role of negative feedback and found that when teachers explicitly mentioned the fact that an error was made in their negative feedback, it led to a higher learning gain than when they did not, which suggests that explicit negative (or corrective) feedback can be more effective than implicit feedback by using recasts.

Feedback is not always negative or corrective, it can also be positive. In general, teachers mostly use positive feedback explicitly (praise) and not implicitly (Hattie & Timperley, 2007). The advantage of praise is that it approves children's answers and makes the task encouraging and motivating (Henderlong & Lepper, 2002). When children receive positive feedback, they become happy, and are therefore more committed and intrinsically motivated to complete a task. However, there are also downsides to providing positive feedback. When children receive too much positive feedback, they rely on the feedback and will not learn when they do not receive the feedback anymore (Henderlong & Lepper, 2002). In addition, when the use of praise is nonspecific or ambiguous, such as saying "good job" or "beautiful" makes children not understand what part of their answer elicited the feedback and they will not know how to respond (Hamilton & Gordon, 1978). Thus, positive feedback should refer to the learning task and at the same time remain motivating enough in order to be effective.

Engagement seems to have a positive effect on language learning (Christenson et al., 2012). A considerable amount of studies have shown that robots are engaging interaction partners for both adults and children(see for an overview Kanero et al., 2018). Engagement normally starts high due to the novelty effect but then seems to decrease over time (Kanda et al., 2007; Rintjema et al., 2018; Kory-Westlund & Breazeal, 2015). When talking about engagement, it can be helpful to distinguish two kinds of engagement: robot engagement, referring to how engaged a child is with the robot, and task engagement, which focuses on how engaged a child is with the learning task. Clearly, these are not necessarily the same: a child can be very engaged with their social partner, the robot, but not with the task, or vice versa. Moreover, the effect of these different engagement types on learning gain can differ. For example, one study by Kennedy et al. (2015) used a highly engaging robot partner and, as a result, children were so distracted by the robot that they focused less on the task and therefore learned less. In their study, children who were highly engaged with the robot, learned less instead of more while it is possible that children who are highly engaged with the task, will still learn more. Consequently, it is useful to measure both types of engagement: task engagement and robot engagement.

Research in HRI has looked at many ways of keeping general engagement high, but did not investigate the role that different types of feedback could play here. For example, Ahmad et al. (2019) looked at the role of adaptive feedback on the children's emotion on engagement, but they did not investigate the effect of different types of feedback.

Feedback, however, can have an influence on children's motivation and their self-evaluation (Zentall & Morris, 2010), which –in turn– can influence engagement. Blumenfeld et al. (2006) suggested a feedback loop: in order to increase children's engagement, children first have to be motivated, which will then increase their interest in the task, which in turn will engage children followed by the children's learning gain. When children improve their language skills, this can lead to even higher motivation and further result in a higher engagement.

The influence of feedback on motivation depends on the type of feedback. For instance, praise that is specifically linked with the children's effort (e.g., "You are a good drawer" after drawing a picture) motivates children more than other types of praise, even when only 75% of the praise is linked with effort (Zentall & Morris, 2010). Moreover, Corpus & Lepper (2007) showed that for preschool children all praise enhanced motivation when they compared it with neutral feedback ("OK"). They compared motivation of preschool children with older children, and found that only for older children (fourth and fifth graders) the type of praise had an influence on their motivation, while preschool children benefited from all feedback equally. Another study found similar results: Morris & Zentall (2014) measured ambiguous praise ("Well done!", "Yeah", "Awesome") and found higher persistence, higher self-evaluations and fewer fixations on later mistakes. Apparently, children interpret ambiguous praise in the most beneficial manner for themselves. However, they also found that the use of gestures ("Thumbs up" and "High five") resulted in the highest self-evaluations.

The reason why feedback has an influence on motivation and, therefore, engagement can be explained by the Self-Determination Theory (Deci & Ryan, 1985). This theory poses that learners continue a task longer when their motivation is based on intrinsic aspects, such as pleasure and satisfaction, compared to when motivation is based on external rewards (Deci & Ryan, 1985). This intrinsic motivation arises particularly when a task contains autonomy and competence and is strengthened by a sense of relatedness between learner and teacher (Ryan & Deci, 2000). For example, autonomy increases when a learner can choose themselves what kind of activity to do, or when he or she receives informative rewards and non-controlling instructions. A higher degree of autonomy leads to increased intrinsic motivation and, in turn, higher levels of engagement. Moreover, competence increases with praise (Blanck et al., 1984), because it enhances the children's feeling of being capable to successfully complete a challenging task. Competence, especially in combination with autonomy, plays a considerable role in retaining intrinsic motivation. There are also disadvantages of praise, for example, when children first receive praise but are not able to successfully complete the task, their motivation can decrease (Zentall & Morris, 2012). Moreover, too much positive feedback can decrease the children's own curiosity (Henderlong & Lepper, 2002).

Negative feedback has been found to decrease intrinsic motivation, specifically the feeling of competence (Deci et al., 1991). It can potentially decrease children's self-efficacy or their active participation and engagement in the learning task, because children become unmotivated when receiving negative feedback (Wojitas, 1998). On the other hand, negative feedback can also have a positive influence on motivation, as it can help children to understand what they are trying to learn and to correct their mistakes (Hattie & Timperley, 2007). Kluger & deNisi (1996) suggest that, similar as with praise, the effect of feedback is not only dependent on a link between behavior and feedback, but also on how the feedback was provided and how the learner interprets the feedback.

The combination of praise and negative feedback can be challenging enough for children, but at the same time can motivate children enough to want to continue with the task.

For example, if children additionally receive negative feedback to correct their mistakes and hear praise when they correctly answer a question, this can enhance the effect of both feedback types. Summarizing, feedback has the potential to both engage and disengage children (Dempsey & Sales, 1993), depending on the type of feedback given. Feedback (especially praise) can increase the intrinsic motivation of children, which increases their engagement. Engaged children are more motivated, learn faster, will be more likely to complete the task and to repeat the task, which leads to a better result (Dörnyei, 1998). However, it is not clear yet whether the rules that apply to human teacher-child interactions also apply to robot-child interactions.

Feedback in child-robot interaction

Studies with educational robots for children that have explicitly looked at the role of feedback are sparse. While many studies have incorporated the use of feedback, specifically praise (Kennedy et al., 2016; Gordon et al., 2016; Mazzoni & Benvenuti, 2015; Kory-Westlund & Breazeal, 2015), they did not test the effect of feedback on the children's engagement or learning gain nor the effects that different forms of feedback may have. These studies investigated the role of praise either by incorporating it as part of a robot's strategy (Kennedy et al., 2016; Kory-Westlund & Breazeal, 2015), by looking at specific responses of children on occurrences of praise (Serholt & Barendregt, 2016) or on the effect of timing of the praise (H. W. Park et al., 2017). It seems that children notice the praise and react to it, however, these studies did not investigate its direct effect on engagement and learning gain. For example, Kennedy et al. (2016) compared a high verbal availability robot and a low verbal availability robot. The high verbal availability robot used –among other social behaviors– more expressive praise than the low verbal availability robot. Children of approximately 8 years old practiced different French grammar rules with one of the robots. The authors found no significant difference in learning gain for the robot that used more expressive positive feedback, but the children reported to have noticed the praise and payed attention to it.

In another study, Serholt & Barendregt (2016) investigated children's responses to the robot's praise. In their long-term study, the robot gave praise on the children's performance of the previous session. Positive feedback did not go unnoticed, 70% of the children acknowledged the robot during feedback through verbal or gestural responses such as smiling. Similarly, H. W. Park et al. (2017) explored whether the timing of a robot's praises would influence the engagement of children. Children had to tell a robot a story and the robot reacted on their emotional level as a form of feedback. For example, when children had a high energy level, the robot played a large excited motion. Park and colleagues compared two conditions, one with a robot that reacted every 5 seconds on the child without changing its energy level, and one with a robot that reacted when the child stopped talking and changed the energy level of its responses appropriately. The children seemed to be more engaged with latter robot that changed its feedback to their energy level. Likewise, Kory-Westlund & Breazeal (2015) used a non-humanoid robot to teach children a second language and found that children learned with a social robot more than with a non-social robot. Both robots used positive phrases when children were correct, e.g., "Good job!" or "You're working hard!" and only provided hints with an incorrect answer, e.g., "I think it was that one". However, the social robot added expressive phrases based on the child's emotional state (e.g., when children were excited, the robot first reacted with "woo hoo" before the feedback).

While many robots use praise, which is an explicit form of positive feedback, explicit negative feedback is not often used in child-robot studies. Typically, studies incorporated implicit feedback by using hints (e.g., "I think it was the other one", Gordon et al., 2016) or by introducing doubts ("Are you sure?", Mazzoni & Benvenuti, 2015).

Three studies that specifically investigated the effect that feedback has on learning and/or engagement are those by Resing et al. (2019), Hindriks & Liebens (2019) and Ahmad et al. (2019).

Resing et al. (2019) reported a study where 6 till 9-year-old children had to solve a puzzle together with an owl-like robot that either helped them by giving feedback or did not provide any help. The help-providing robot used both verbal and non-verbal feedback. It shook its head and had blinking eyes when their answer was incorrect as a way of providing non-verbal (explicit) negative feedback, or nodded and said "Well done!", with (different) blinking eyes as a form of explicit positive feedback. Children trained by the robot with feedback became better in solving new puzzles than children trained with the other robot. However, children showed large individual differences in the number of corrections they needed.

Hindriks & Liebens (2019) conducted a between-subject study with 7-9-year-old children who had to solve mathematical problems. They compared a robot providing feedback designed to show the children which specific error they made (e.g., forgetting to add one number) with a robot asking the children to think aloud when solving the problem. This error-specific feedback did not affect children's learning gain. Instead, children, showed, again, individual differences. Children who showed to have difficulties with the math problems, appreciated the feedback more than children who did not have problems with math.

Ahmad et al. (2019) addressed individual differences between children and compared in a between-subjects design a robot that adapted its feedback with one that did not. They studied how children between 10 and 12 years old responded to the robot's feedback during two weeks. The robot adapted its feedback behavior to the children's emotional state. For example, when children were rated as happy the robot used that in its feedback ("You are looking happy, and I'm happy that you are in front of me. Let's learn another word"). During the game, the robot kept referring to the game outcome, only in the post-test the robot provided feedback on learning performance ("I am happy that you got it wrong in session one, but this time your answer is correct" or "It's sad that you didn't remember this word, the correct answer is ..."). Ahmad and colleagues found that the children's engagement remained relatively high (or stable) when interacting with the adaptive robot, while their engagement lowered over time with the non-adaptive one. Moreover, children's learning gain was higher with the adaptive robot, compared to the non-adaptive one. While these results are promising, this study did not investigate the effect of different feedback strategies.

Generally, developers of robot tutors base the educational strategies of the robot on the already existing human studies and use those strategies in their child-robot interactions without studying whether these strategies are similarly effective. Most child-robot studies use praise as a motivator in their experiments and are hesitant to use explicit negative feedback. It is not clear what type of negative feedback works best for robots, although in educational studies it seems that mentioning the children's mistake seems to be more effective for language learning. In this chapter, we address this gap in knowledge by investigating the effect of different forms of feedback on both task engagement, robot engagement and learning gain.

3.2.2 Teachers' feedback

In preparation of this chapter, we carried out a survey among student teachers concerning their views on how a robot should provide feedback. The aim of this survey was twofold: 1) To gain insights how student teachers would think the robot should provide feedback to children giving correct and incorrect answers in a tutoring setting, and with varying levels of the children's engagement at the time feedback is given. 2) To create a data set with different feedback phrases that student teachers would use. We interviewed student teachers instead of practicing teachers, because students are more likely to work with technologies in the future, such as social robots, than teachers who already worked for many years. Moreover, receiving many responses was more feasible with student teachers than with teachers.

In our survey, we showed 27 student teachers 40 video fragments of both engaged and disengaged children interacting with a robot in a L2 tutoring experiment reported in de Wit et al. (2018). All fragments showed a robot teaching 5- to 6-year-old Dutch children animal words in English as a second language. In each fragment, the robot expressed an English word and asked the child to select –on a tablet– the animal he or she thought that the word referred to. The fragment ended right after the child answered to this request. After watching each fragment, the student teachers were asked to provide a feedback suggestion. The survey was carried out in a between-subject design with two conditions: in one condition (closed questions), student teachers could choose between six feedback strategies (three positive and three negative), and in the other condition (open-ended questions) they could freely write the feedback themselves. This closed questions survey would provide insights of what strategy student teachers would choose, and the open questionnaire would create a data set of different feedback phrases.

We did not find a difference between student teachers' suggestions for engaged or disengaged children. However, we found that the suggested forms of feedback differed substantially between the closed and open-ended questionnaires: In the closed questions survey, the majority of the student teachers chose to use an explicit positive phrasing together with an explanation in the form of a translation ("Goed zo! Een 'hippo' is een nijlpaard" (Dutch)- "Well done! A 'hippo' is a hippo" (English)), and they chose a correction of the child's answer through repetition and translation of the target words ("Een hippo is een nijlpaard, je moet de nijlpaard aanraken" (Dutch) - "A 'hippo' is a hippo, you have to touch the hippo" (English)) as a means of providing implicit negative feedback.

In the case of the open-ended survey, the student teachers chose for both positive and negative feedback to only provide an explicit phrasing without repeating the target words for both positive feedback ("Goedzo" (Dutch) - "Well done" (English)) and negative feedback ("Helaas dat was niet goed" (Dutch) - "Unfortunately, that was not correct" (English)). Moreover, we found that in the open-ended questionnaire student teachers varied their phrasing of the feedback considerably.

After the surveys were analyzed, we discussed the findings with a subset of the student teachers. They suggested two main reasons why these results differed. Firstly, correction and explanation (e.g., through repetition of target words) is essential for negative feedback. This was the main reason why they chose to repeat the target words in the closed-ended questionnaire. Secondly, they indicated that variation in the form by which feedback is provided is also crucial. The robot should not repeat the same phrase throughout the whole session. Student teachers participating in the open-ended questionnaire focused more on creating varying feedback phrases and less on the repetition of the target word.

Based on these findings, we concluded that the *preferred* feedback strategy would combine the results from the closed questions survey with the open-ended survey: take an explicit feedback phrase (e.g., "Well done" or "That's wrong"), add a repetition of the target word, and provide children an extra attempt when their answers are incorrect. Since variation is key, the feedback phrases should vary, based on the data set created by the open-ended survey.

3.2.3 This study

This study investigated whether 5- and 6-year-old children are more engaged with the task and with the robot, and learn more words when participating in an L2 training with a robot that provides feedback as recommended by the student teachers (preferred feedback), compared to a robot that provides feedback contrary to what was recommended by the student teachers (dispreferred feedback), and compared to a robot that provides no feedback at all (no feedback). As our survey with student teachers revealed, providing adequate feedback is a complex matter that consists of multiple strategies, which are hard to separate, thus making it difficult to investigate such individual factors experimentally. We, therefore, decided to combine multiple factors in our preferred and dispreferred feedback strategies, and explored to what extent these strategies, as performed by a robot, influence children's engagement and learning gain in an L2 tutoring scenario.

Every child received three sessions with different robots, each providing a different form of feedback, thus allowing us to investigate how children react to the different forms of feedback using a within-subjects design. We based the training sessions on previous studies in which children played an "I spy with my little eye" game with a NAO robot to learn different L2 words (Schodde et al., 2019; de Wit et al., 2018).

Based on previous findings in literature regarding the role of feedback in L2 learning, and previous studies that addressed feedback in child-robot interactions (Ahmad et al., 2019), we hypothesized that children would be more task-engaged and robot-engaged when receiving (either preferred or dispreferred) feedback than when they did not receive feedback (H1a). Especially positive feedback was expected to increase the children's intrinsic motivation for the task and thus their engagement. We also hypothesized that children would remember more words when receiving feedback than when receiving no feedback (H1b). Feedback can help to understand whether an answer is correct or not and may indicate what the correct form should be, thus providing insight into the learning process and helps to improve the learning performance.

Moreover, we hypothesized that children would be more task-engaged and robot-engaged with (H2a) and would remember more words from (H2b) a robot that provides feedback as preferred by a student teacher compared to a robot that provides dispreferred feedback. When feedback is varied (as in the preferred feedback strategy), children were expected to pay more attention to it, boosting their confidence and with that their task engagement. The varied feedback of the robot could additionally increase the children's interest in the robot and with that their robot engagement. In contrast, when a robot repeatedly used the same phrase as feedback (dispreferred feedback), children might have gotten tired of this repetition and as a result would pay less attention to the robot. Additionally, children could practice with the preferred feedback once more in the case of a mistake and thus improved their knowledge, which they could not with the dispreferred feedback strategy and which might have lead to an increase in their task engagement. Finally, the preferred feedback also provided children with an explicit notion where the mistake was been made, what went wrong and how they could fix it by trying again (the three rules of good feedback according to Hattie & Timperley, 2007).

3.3 Method

The research questions, hypotheses and analyses in this study have been preregistered at AsPredicted[1] and the source code has been made publicly available[2].

3.3.1 Design

The study was a within-subjects design, where all participants were assigned to all feedback strategies/conditions (each session a different strategy). The strategies for providing feedback were based on the survey asking student teachers how they would make the robot provide feedback in situations comparable to the ones in this experiment, translating to a preferred strategy and dispreferred strategy. The order of the feedback strategies and word sets were counterbalanced using a 3x3 latin-square to reduce an order effect. The three strategies/conditions were

1. Preferred feedback
2. Dispreferred feedback
3. No feedback

[1] `https://aspredicted.org/qg6dx.pdf`
[2] `https://github.com/l2tor/feedback-study/`

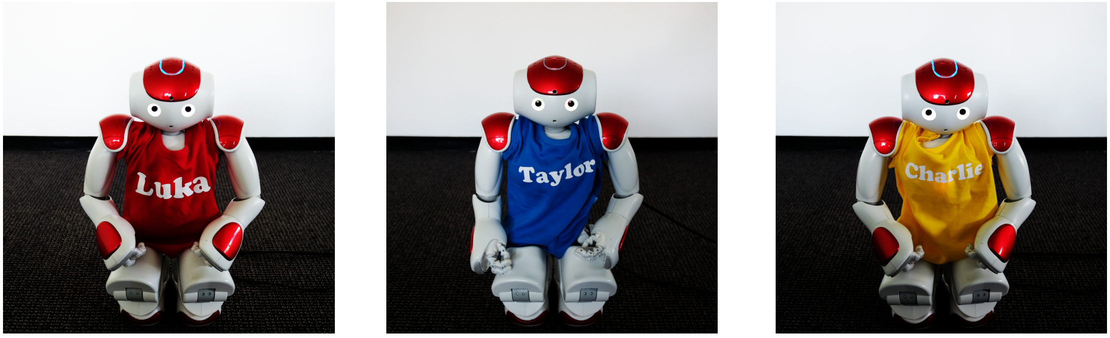
Each child received three sessions with the robot, and could learn 18 words in total and 6 in each session. In all conditions, all sessions were the same, except for the words learned, the feedback strategy that the robot used, the name and the shirt the robot was wearing (to give the impression that children were playing with three different robots, see Figure 3.1).

3.3.2 Participants

In total, 72 native Dutch-speaking children aged five and six years participated in the current study. The participants were recruited from three elementary schools located in the Netherlands. Bilingual children were excluded from the study. A pre-test showed that 12 children were familiar with more than half of the target words and these children were excluded from the study in accordance with the exclusion criteria of our preregistration. Furthermore, four children dropped out of the study for various reasons like unwillingness to continue (3) or sickness (1). This resulted in 56 children (M_{age} = 5 years and 6 months, SD = 5 months, 28 boys and 28 girls) participating in the final experiment. All parents gave informed consent for the participation of their child. The study was conducted in accordance with the Declaration of Helsinki, and received ethical approval from the Research Ethics committee of Tilburg School of Humanities and Digital Sciences.

3.3.3 Materials

The Softbank Robotics NAO robot and a Microsoft Surface tablet computer were used. The sessions involved one-on-one interactions between robot and child. We did not rely on automatic speech recognition because speech recognition has been shown to not work well with this age group (Kennedy et al., 2017). Instead the experimenter used a Wizard of Oz technique when the child had to say something to the robot in the beginning of the experiment. The robot was placed in a crouching position in an angle of 90 degrees next to the sitting child to give the robot the same perspective of the child, while still being able to face the child. The tablet was placed on top of a small box in front of the robot and child. A video camera placed on a tripod facing the child to record the child's responses and facial expressions. A second camera was placed from the side to get a more complete overview of the interactions. Each session was distinguished by a different color shirt and robot name (see Figure 3.1). We used the different shirts and names to make it known to children that they would play with three different robots, with different robot behaviors (namely the robot feedback strategies). The shirts and robot names were not linked to feedback conditions or different word sets, but rather to the session number. In other words, all children started with the robot wearing the red shirt called Luka during the first session and ended with the robot wearing the yellow shirt called Charlie.



a: Robot first session b: Robot second session c: Robot third session

Figure 3.1: (a,b,c) show the different shirts for each sessions. All children saw the robot wearing the red shirt during the first session and all children saw the robot wearing the yellow shirt during the last session.

Target words

In total 18 target words were selected and during each session, children learned six target words. Target words were selected such that children can be expected to have acquired those in their first language but arguably not in their second language. Moreover, we selected words that would not be too similar in their L1 and in their L2 (e.g. not "Olifant" (Dutch) and "Elephant" (English)). All 18 words were divided in three word sets based on their frequency in the children's first language. We used the dataset of (Schrooten & Vermeer, 1994) and placed each word in a frequency bin. Words in the same bin were randomly assigned to the different word sets. For example, the word "dog" was from the same frequency bin as the words "bird" and "horse" and were thus added to different word sets. Table 3.1 contains an overview of all target words with their frequency. We used cartoon-like images of the target animals during the experiment (see Figure 3.2 for examples).

Pre-test

Before the children started the three sessions, we tested their L2 knowledge of the 18 target words with a comprehension test which was a picture-selection task. In this test, children

Table 3.1: Target words with their frequency scores in Dutch taken from Schrooten & Vermeer (1994). Words that have a higher score are more familiar to children in Dutch.

Word set 1			Word set 2			Word set 3		
Dutch	English	Freq	Dutch	English	Freq	Dutch	English	Freq
Hond	Dog	98	Vogel	Bird	72	Paard	Horse	64
Kikker	Frog	27	Kip	Chicken	30	Konijn	Rabbit	48
Vlinder	Butterfly	22	Nijlpaard	Hippo	16	Varken	Pig	36
Papagaai	Parrot	9	Slang	Snake	14	Eekhoorn	Squirrel	13
Haai	Shark	9	Slak	Snail	14	Zeehond	Seal	10
Neushoorn	Rhino	9	Walvis	Whale	9	Hert	Deer	9

were presented with a pre-recorded target word spoken by a bilingual speaker of Dutch and English and asked to choose which one out of four pictures matched this word ("Waar zie je een dog?" (Dutch) "Where do you see a: dog?" (English)). The presentation of the target words in the pre-test was randomized for each child. We presented each target word one time during the pre-test.

Post-test

The children's long-term knowledge was tested between two and three weeks after the last session with the comprehension test. The test was the same as the pre-test only this time, each target word was presented three times in a random order to reduce chance level performance due to guessing. The reason for not doing so in the pretest was to reduce the chance of children learning from this task (Smith & Yu, 2008). A word was registered as correct if it was selected correctly at least twice out of the three trials. Additionally, we tested three different pictures of the animals in order to generalize the children's knowledge. To be more specific, we used a cartoon-like picture, a drawn picture (the same as in the experiment) and a photograph of the target animal.

In addition to the measurements described in this chapter we also carried out a perception questionnaire of the robot at the end of all sessions. We will not discuss those results because this questionnaire is beyond the scope of this chapter.

3.3.4 Tutoring sessions

The sessions were based on the children's game "I spy with my little eye" and on the interaction described in Schodde et al. (2019). The whole interaction was in the children's L1, except for the target words. Before the three tutoring sessions, children had a group introduction to the robot and took a pre-test. The tutoring session had four parts which were all repeated during all three tutoring sessions:

1. Start phase. The robot explained that he was a friend of the group introductory robot, he asked for the child's name, age and some questions about their favorite animals and games. The robot finished with saying that "I spy with my little eye" is his favorite game and that he wants to play that with the children. He then explained the rules of the game.

2. Concept binding of the target words. To teach children the target words, the tablet showed an animal on the screen, the robot said the L2 word with the L1 translation and asked the child to repeat the word (e.g. "Een vogel is een bird in het engels, zeg mij maar na bird" (Dutch). "A bird is a *bird* in English, repeat after me *bird*" (English)). Only after the child had repeated the animal, they continued to the next animal. When a child did not repeat the robot, the experimenter asked the child to listen to the robot and repeat after the robot. If a child was very hesitant to repeat the word, the experimenter would say it together with the child.

3. Training rounds. After the concept binding the robot explained to the child that he would ask for an animal and that the child had to search for it on the tablet screen. They first practiced with an L1 word that was no target ("Ik zie, ik zie wat jij niet ziet en het is een eenhoorn, zoek maar naar de eenhoorn", "I spy with my little eye a unicorn, please search for the unicorn"). For each target word the tablet showed the target animal with three distractors (see Figure 3.2a). After the L1 practice round, the robot and child also practised once in L2. After these two practice rounds they started the training of the target words. The robot constantly asked the child to search for a target word ("Ik zie, ik zie wat jij niet ziet en het is een <target word> zoek maar naar de <target word>", "I spy with my little eye a <target word>, please search for the <target word >"). Depending on the condition the robot provided feedback or not and the child continued to the next animal. There were 24 rounds in total, each animal was trained four times, which made the L2 exposure to all animals ten times in total for all conditions (twice in the concept binding, eight times during the practice rounds).

4. In-game test. After each session there was an in-game test that tested the short-term

memory of the target words. The tablet screen showed all animals of that tutoring sessions and a bucket of grapes (see Figure 3.2b). Each round, the robot named an animal and the child had to feed this animal with one of the grapes. The robot asked the animals in random order and after each round the order of presenting the animals on the screen was shuffled.

All conditions had the exact same design, meaning that the session structure was the same, the tablet output was the same and the behavior of the robot was the same, except for the feedback. In all conditions, the robot used the standard following-gaze feature of NAO.

a: Training round   b: In-game test   c: Second attempt

Figure 3.2: (a) Training rounds. Each round the robot asked for one animal (b) In-game test. Children had to drag a grape to the animal that the robot named (c) Second attempt after wrong answer. Children were allowed to correct themselves in the preferred feedback condition. In this example, the child wrongly chose a butterfly instead of a parrot and could correct his/her mistake by selecting the correct one.

3.3.5 Feedback conditions

All feedback was provided in the children's L1 to keep the L2 exposure consistent between conditions. A comparison of the different types of feedback can be found in Table 3.2. The feedback conditions were based on the student teachers' preferred response for the robot (preferred feedback), the opposite (dispreferred feedback) and a control condition was added where the robot did not use any feedback. Preferred and dispreferred feedback different on multiple aspects:

1. Variation. The robot used a variety of positive and negative feedback in the preferred feedback condition and no variation in the dispreferred feedback condition. We based the phrases on the student teachers' open-ended survey and can be found in Table 3.3. The robot randomly chose between six verbal phrases for positive feedback and neg-

ative feedback and the same phrase was never used twice in a row. We only added variation to the preferred strategy because the student teachers considered this crucial.

2. Extra attempt. The robot let children to try again after an incorrect answer in the preferred feedback condition and not in the other conditions. This was based on the student teachers' closed-ended answers where they relied heavily on the answer with the extra attempt. During the extra attempt, the tablet would only display the correct target word and the children's incorrect answer to help the children distinguish the two answers (see Figure 3.2c). After children correctly answered their second attempt, they received positive feedback.

3. Repetition. In the preferred condition, the robot would repeat the target word, either in addition to positive feedback or in addition to noting the mistake including the child's wrong answer. However, this was only done in 50% of all feedback to reduce the amount of repetition and because the student teachers did not always use a repetition (only in the closed-ended questionnaire and not in the open-ended questionnaire). The robot would only repeat the target word in the children's L1 (i.e. Dutch) to keep the amount of L2 exposure consistent over all children and to only focus on the effect of feedback.

4. Non-verbal feedback behavior. The robot used some non-verbal behavior when the child was correct in the preferred feedback condition, but not in the dispreferred feedback condition. This non-verbal behavior consisted of the robot nodding and displaying a rainbow coloured pattern in the LED-eyes to indicate happiness.

After the feedback was provided (or after the child's answer in the no feedback condition), the game continued to the next target word.

3.3.6 Procedure

Robot introduction and pre-test. One week before the experiment, the children participated in a group introduction to familiarize themselves with the robot. During this introduction, based on Vogt et al. (2017), children learned how the robot moves and how to talk to it, and they played a game where they had to imitate the robot and they danced together. Unlike the robots during the experiment, this robot was not wearing a shirt. After this group introduction the children carried out a pre-test on their prior English knowledge in one-on-one sessions, as explained in Section 3.3.3.

Table 3.2: An example of the robot's feedback in the different feedback conditions.

Condition	Correct answer	
	Dutch	English
Preferred	Goed gedaan, het was een vogel.	Well done, it was a bird.
Dispreferred	Goed gedaan.	Well done.
No feedback	-	-
	Incorrect answer	
	Dutch	English
Preferred	Helaas, je hebt een vogel aangeraakt. Laten we het nog eens proberen!	Unfortunately, you selected a bird. Let's try again!
Dispreferred	Helaas, dat is niet goed.	Unfortunately, that was not correct.
No feedback	-	-

Experiment. At least one week after this group introduction and the pre-test, we started the first tutoring sessions with the children. Children participated in a quiet room away from their classrooms. After the child was collected from her or his classroom for the first session, he or she was told that he or she would play a game on a tablet with a friend of the introduction robot. This was repeated every new session so each child saw four "different" robots in total (one introduction robot and three robots in the tutoring sessions). When the child entered the room with the robot, the experimenter told the child to sit in front of the tablet next to the robot and started the experiment. After the child finished the 24 rounds of "I spy with my little eye" and the subsequent in-game post-test, the experimenter filled in a questionnaire with the child about the robot. When this questionnaire was completed the experimenter brought the child back to the classroom. This was repeated for three times with at least one day in between the different sessions.

The interaction was semi autonomous, except for the detection of children's speech when they repeated the target words as instructed. For detecting the child's speech, the experimenter would press a button on a control panel once the child had repeated the robot's utterance. The interaction was a one-on-one interaction, but the experimenter stayed in the same room to intervene when necessary. For example, when a child did not repeat after the robot, the experimenter would try to encourage the child to repeat after the robot. Moreover, when the child had a question, the experimenter would say that she did not know the answer

Table 3.3: The preferred feedback utterances. The robot's feedback varied between six different options.

Positive	
Dutch	English
Goed gedaan!	Well done!
Knap hoor.	Impressive.
Ja goed gedaan!	Yes, well done!
Ga zo door!	Keep going!
Super!	Great!
Heel knap gedaan.	Really impressive.
Negative	
Dutch	English
Helaas dat was niet goed.	Unfortunately, that was not correct.
Sorry deze is niet goed.	Sorry but this is not correct.
Helaas, probeer het nog een keer.	Unfortunately, try again.
Jammer, we proberen het nog eens.	What a pity, let's try again.
Ah jammer, denk nog even goed na.	Ah pity, think again.
Super goed geluisterd, maar dat was niet goed, probeer het nog eens.	You listened very well, but this was not correct, try again.

and directed the child's attention back to the robot. In other cases, when a child had to go to the bathroom, the experimenter paused the experiment and walked with the child to the bathroom and back. The duration of each session was around 11 minutes (Preferred: M = 14 minutes, SD = 2 minutes, Dispreferred: M = 11 minutes, SD = 1.5 minutes, No feedback: M = 10 minutes, SD = 1 minute).

Post-test. Two weeks after the last session, the children were collected from the classroom once more for the post-test.

3.3.7 Engagement coding and analyses

Engagement coding

Engagement was determined by manual coding of half of the data. Before coding, the two raters followed a coding training and practiced with different videos. Each video was rated on a Likert scale from 1 to 5, with 1 being a low level of engagement and 5 being highly engaged.

We measured *task engagement* that includes the attention that the child payed to the robot as instructor, but also to the task displayed on the tablet screen. Children were fully engaged, when they were completely 'absorbed' in the activity, were open for new information, were very motivated, enjoyed the task and wanted to play with the robot (Laevers, 2005). Additionally, we rated *robot engagement* that measures the children's attention and interest at the robot as a social interaction partner. Children were fully engaged with the robot, when they were interacting with the robot as a social conversation partner.

The coding scheme was based on the ZIKO coding scheme (Laevers, 2005). The ZIKO scheme describes a measurement for, among others, children's engagement. It is designed for child-task engagement in open classroom settings. We adapted the scheme to include specific cues for this experiment by including cues such as, attention toward the experiment leader instead of the robot or tablet and child is randomly clicking on the tablet in order to continue.

Each engagement level had specific cues for the rater to look for. For example, children scored high on *task engagement* when they were not only looking at the task and robot, but also actively searching for the different animals on the tablet and were fully committed to the task. In contrast, when children turned away from the robot and task, did not perform anything related to the task and were fiddling, this resulted in a low engagement. Children who played the game but did not pay all their attention to it received an average task engagement rating. In the case of *robot engagement* we added social engagement cues, such as looking at the robot, having spontaneous conversations with the robot, but it also included the children's posture toward the robot (a closed posture indicating a low robot engagement and an open posture indicating a high robot engagement). For all specific cues and information see the coding scheme in Chapter 4 and on Github[3].

For the engagement coding, we pseudo-randomly selected half of the children, excluding children who took a break during the interaction (e.g., when they had to go to the bathroom), which happened in 11 cases. Twenty percent of the selected data was coded by two raters and their inter-rater agreement was considered moderate to good using the intraclass correlation coefficient (ICC_{task} = .70, 95% CI[.37, .76], ICC_{robot} = .80, 95% CI[.62, .90]) (Koo & Li, 2016). For analyses, we only used the data of the first rater. We extracted two two-minute video fragments of the interaction: one at the beginning of the training rounds during the interaction and one at the end of the interaction.

The engagement rating of both fragments were combined to get a more reliable measure of the child's overall engagement during the session. This resulted in 210 engagement ratings

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

in total.

Analyses

To investigate the effect of the different feedback strategies on children's engagement, we conducted a repeated measures ANOVA with the feedback strategy as the independent variable (three levels) and engagement as a dependent variable.

In addition, to investigate the effect of the feedback strategies on learning gain, we carried out a two-way repeated measures ANOVA with the children's scores as a dependent variable and two strategies: (1) feedback strategy (three levels) and (2) test moment (the pre-test and the delayed post-test).

Using planned contrasts, we compared the effect of preferred and dispreferred feedback with no feedback on engagement and learning gain for H1 and preferred feedback and dispreferred feedback for H2. Moreover, to investigate the effect of the feedback strategies on short-term learning gain, a one-way repeated measures ANOVA with feedback strategy as the independent variable and the results of the in-game test as the dependent variable was performed.

3.4 Results

We have made the data set for this experiment publicly available[4]. In this section we report the children's engagement and their learning gain during the sessions. In addition, we report on the possible relation between learning gain and the children's engagement. Children received positive feedback during all 24 rounds in the preferred feedback condition and on average 14.30 times during the dispreferred feedback condition.

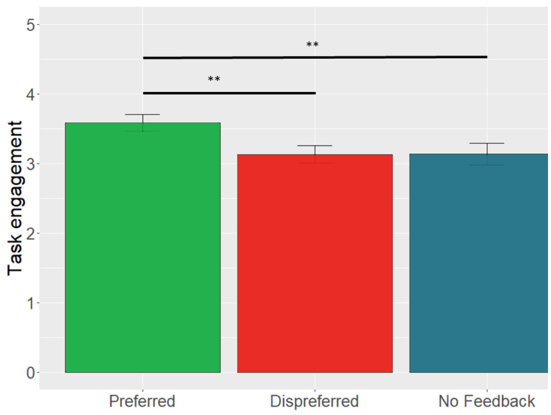
3.4.1 Engagement

Table 3.4 shows the overall results of both engagement types for the different sessions and different conditions. Overall, task engagement ($M = 3.28$, $SD = 0.81$) was slightly higher than robot engagement ($M = 3.06$, $SD = 0.92$). The two engagement types were moderately correlated ($r(105) = .50, p < .01$), indicating that they both measure a different type of engagement.

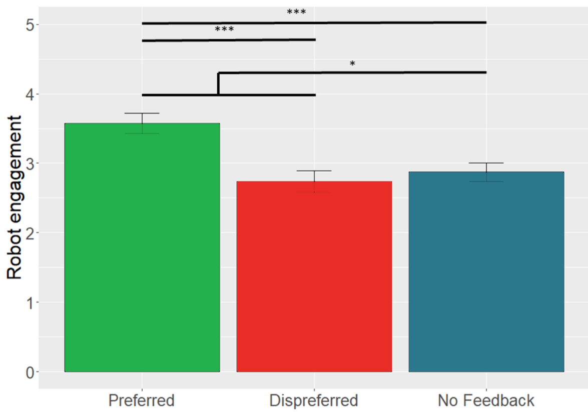
[4]https://doi.org/10.34894/ZEIKLY

Table 3.4: Average task engagement and robot engagement rating over time (SD).

Feedback strategy	All sessions		Lesson 1		Lesson 2		Lesson 3	
	Task	Robot	Task	Robot	Task	Robot	Task	Robot
Preferred	3.59 (0.71)	3.57 (0.87)	3.88 (0.63)	3.92 (0.85)	3.58 (0.86)	3.65 (0.80)	3.27 (0.58)	3.12 (0.81)
Dispreferred	3.13 (0.74)	2.74 (0.90)	3.03 (0.49)	2.50 (0.63)	3.09 (0.79)	3.09 (1.03)	3.23 (0.84)	2.50 (0.83)
No feedback	3.14 (0.91)	2.87 (0.79)	3.50 (0.91)	3.11 (0.88)	3.20 (0.69)	2.70 (0.60)	2.55 (0.90)	2.72 (0.84)
Overall	3.28 (0.81)	3.06 (0.92)	3.54 (0.78)	3.27 (0.97)	3.26 (0.79)	3.13 (0.90)	3.05 (0.82)	2.78 (0.84)



a: Task engagement



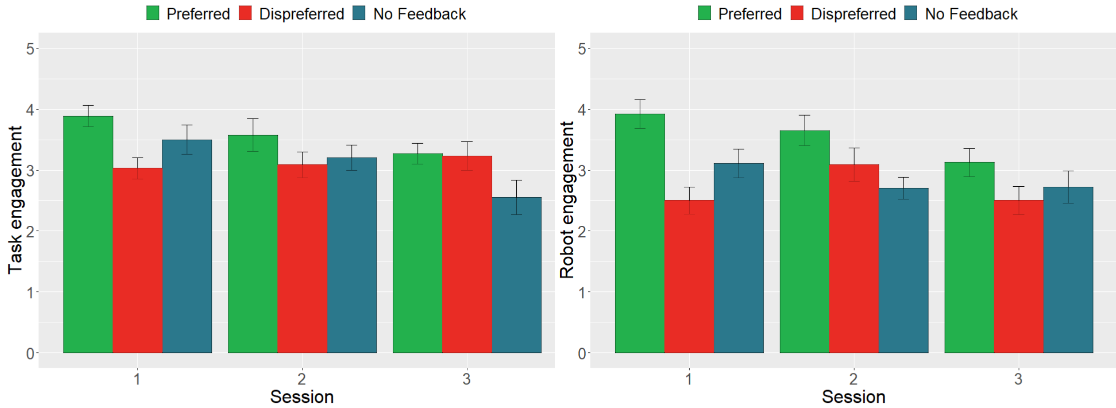
b: Robot engagement

Figure 3.3: Average engagement ratings per condition. Error bars show 95% confidence interval. *p<.05, **p < .01, ***[< .001

Task engagement

Contrary to our expectations, planned contrast analyses for comparing both preferred feedback and dispreferred feedback combined ($M = 3.36, SD = 0.76$) with no feedback ($M = 3.14, SD = 0.91$) showed no significant difference in task engagement ($F(1, 34) = 3.96, p = .06, \eta_p^2 = .10$). However, as Figure 3.3 shows, children are more engaged with preferred feedback ($M = 3.59, SD = 0.71$) than with dispreferred feedback ($M = 3.13, SD = 0.74$; $F(1, 34) = 13.49, p = .001, \eta_p^2 = .28$). Further analysis using post-hoc comparisons with Bonferroni correction revealed that children were significantly more engaged in the preferred feedback condition than the no feedback condition ($t(34) = 3.26, p = .003$, $M_{diff} = .45$). There was no significant difference between dispreferred and no feedback ($t(34) = -0.06, p = .96, M_{diff} = -0.01$).

Task engagement dropped significantly over time (see Figure 3.4). A repeated measures ANOVA with a Huynh-Feldt correction was performed, because our data violated the assumption of sphericity. The analyses showed that task engagement differed significantly between the sessions ($F(1.64, 55.90) = 7.16, p = .003, \eta_p^2 = .17$). Post hoc tests using the Bonferroni correction revealed that task engagement dropped significantly between session 1 ($M = 3.54, SD = 0.78$) and 2 ($M = 3.26, SD = 0.79$; $t(34) = 2.82, p = .008, M_{diff} = .28$), and session 3 ($M = 3.05, SD = 0.82$; $t(34) = 3.13, p = .004, M_{diff} = .41$) but not between session 2 and 3 ($t(34) = 1.68, p = .102, M_{diff} = 0.21$).



a: Task engagement b: Robot engagement

Figure 3.4: Average task engagement and robot engagement ratings over time and per condition. Error bars show 95% confidence interval. Note that a child who, for example, received preferred feedback in session 1 received different feedback in session 2 and in session 3.

We further tested whether there was an interaction effect between the feedback strategy and the session in which it was used. To this end, we used a mixed ANOVA with order as between factor and feedback strategy as within factor, because this accounts for the order in which participants received the different feedback strategies (for example, it might have had an influence on their task engagement when they received no feedback first and the preferred feedback during the third session). There was a significant interaction effect between order and feedback strategy ($F(10, 58) = 4.43, p < .001, \eta_p^2 = .433$) indicating that the effect of feedback on task engagement varied as a function of when this feedback in the experiment it was administered taking into account that overall task engagement decreased over time. As Table 3.5 illustrates, children's task engagement dropped over time, but not for all orders of the feedback strategies. The task engagement dropped in most situations after children received preferred feedback, task engagement never increased after dispreferred feedback and it either dropped or remained the same for no feedback. An exploratory repeated measures ANOVA on each order indicated that task engagement differed significantly when preferred feedback ($M = 4.0, SD = 0.68$) was provided first, then dispreferred feedback ($M = 2.69, SD = 0.92$) and lastly no feedback ($M = 3.28, SD = 0.81; F(2, 14) = 18.11, p < .001, \eta_p^2 = .72$) and furthermore, when preferred feedback was provided first ($M = 3.70, SD = 0.54$), then no feedback ($M = 2.45, SD = 0.89$) and lastly dispreferred feedback ($M = 3.35, SD = 0.91; F(2, 8) = 8.11, p = .012, \eta_p^2 = .67$). All other orders did not differ significantly (all p values $> .1$).

Table 3.5: The task engagement order effects visualised, a decreasing arrow shows decreasing task engagement and vice versa. P stands for preferred feedback, D for dispreferred feedback and N for no feedback. Task engagement differed significantly for the first two orders with *indicating a p value $< .05$ and **p value $< .001$.

Lesson 1		Lesson 2		Lesson 3
P	$\searrow$	**D**	$\searrow$	**N** **
P	$\searrow$	**N**	$\searrow$	**D** *
D	$\longrightarrow$	P	$\searrow$	N
D	$\longrightarrow$	N	$\longrightarrow$	P
N	$\longrightarrow$	P	$\longrightarrow$	D
N	$\longrightarrow$	D	$\nearrow$	P

Robot engagement

Similarly as for task engagement, we compared the average children's robot engagement score during both the feedback conditions ($M = 3.15, SD = 0.98$) with the no feedback condition ($M = 2.87, SD = 0.79$) using planned contrast analyses. Unlike for task engagement, we found a significant difference in robot engagement between feedback and no feedback ($(F1, 34) = 4.39, p = .04, \eta_p^2 = .11$), albeit with a relatively small effect size. Moreover, children scored higher for robot engagement in the preferred feedback condition ($M = 3.57, SD = 0.87$) than in the dispreferred feedback condition ($M = 2.74, SD = 0.90; F(1, 34) = 43.19, p < .001, \eta_p^2 = .56$). Furthermore, post-hoc comparisons with Bonferroni correction revealed that children were significantly more engagement in the preferred feedback condition than in the no feedback condition ($t(34) = 6.57, p < .001, M_{diff} = 0.70$). There was no significant difference between robot engagement in the dispreferred feedback condition and the no feedback condition ($t(34) = 4.61, p = 1.0, M_{diff} = -.13$).

As Figure 3.4b showed, robot engagement also dropped over time. A repeated measures ANOVA showed a significant difference between the sessions ($F(2, 68) = 4.56, p = .014, \eta_p^2 = .12$). Again, note that the effect size is relatively small. Pairwise comparisons with a Bonferroni correction showed that robot engagement dropped significantly between session 1 and 3 ($t(34) = 2.67, p = .04, M_{diff} = .49$). There was no significant difference between session 1 and session 2 ($t(34) = .87, p = 1, M_{diff} = .14$) nor between session 2 and 3 ($t(34) = 2.27, p = .09, M_{diff} = .35$).

Similarly as with task engagement, we investigated whether there was an interaction effect between the feedback strategy and the session in which the feedback strategy was used. To test this, we used a mixed ANOVA with order as between factor and feedback strategy as within factor. For robot engagement, there was no order effect ($F(10, 58) = 1.58, p = .14$) which indicates that the children's robot engagement was not influenced by different orders of feedback.

3

3.4.2 Learning gain

Children made on average 9.75 mistakes during the 24 rounds (Preferred: $M = 9.95, SD = 5.56$; Dispreferred: $M = 9.30, SD = 5.22$; No feedback: $M = 9.75, SD = 5.41$). Table 3.6 shows the descriptive statistics for the target word knowledge scores for all conditions. Children performed above chance level in the pre-test (chance level = 4.5, $t(55) = 4.27, p < .001, M_{diff} = 1.14$) and post-test (chance level = 2.61, $t(55) = 9.58, p < .001, M_{diff} = 5.25$). As expected, children performed better on the post-test than on the pre-test ($t(55) = -3.88, p < .001, d = .52$), so children clearly learned some vocabulary.

Table 3.6: Average score per condition (SD).

Feedback strategy	Pre-test	Post-test	In-game
Preferred	1.88 (1.38)	2.71 (1.77)	2.80 (1.42)
Dispreferred	1.77 (1.28)	2.59 (1.65)	2.82 (1.62)
No feedback	2.00 (1.31)	2.55 (1.76)	2.75 (1.43)
Total	5.64 (2.00)	7.86 (4.10)	8.38 (3.20)

The two-way repeated measures ANOVA with planned contrasts for both preferred feedback and dispreferred feedback (pre-test: $M = 1.82, SD = 1.33$, post-test: $M = 2.65, SD =$

1.70) showed no difference in learning gain compared to no feedback (pre-test: $M = 2.00, SD = 1.31$, post-test: $M = 2.55, SD = 1.76); F(1, 55) = .47, p = .83$). Furthermore, while children score numerically higher on word knowledge in the preferred feedback condition (pre-test: $M = 1.88, SD = 1.38$, post-test: $M = 2.71, SD = 1.77$) than in the dispreferred (pre-test: $M = 1.77, SD = 1.28$, post-test: $M = 2.59, SD = 1.65$), this difference was not significant ($F(1, 55) = .45, p = .51$).

Table 3.6 also shows the results of the children's in-game tests. Children scored higher than chance in all conditions (chance level = 3, $t(55) = 12.57, p < .001, M_{diff} = 5.38$). Again, feedback strategy did not influence their learning gain, there were no significant differences ($F(2, 110) = .122, p = .89$).

3.4.3 Relation between learning gain and engagement

To investigate whether there was a relation between both engagement types and learning gain, we performed a Pearson correlation analysis and in contrast with what we expected, we found no significant correlation between task engagement and learning gain (preferred: $r(35) = .05, p = .78$, dispreferred: $r(35) = .09, p = .62$, no feedback: $r(35) = .12, p = .50$). Likewise, we did not find a significant correlation between robot engagement and learning gain (preferred: $r(35) = .15, p = .40$, dispreferred: $r(35) = .09, p = .62$, no feedback: $r(35) = .02, p = .90$).

3.5 Discussion

The aim of this chapter was to understand the effects that different types of robot feedback have on children's engagement both with the task, the robot and their learning gain. We derived different types of feedback from a survey with student teachers and implemented them in three different robots, each robot teaching children words from a second language in a single session. One robot provided (teacher) preferred feedback, one provided (teacher) dispreferred feedback, and one provided no feedback at all. All children attended three sessions, each with a different feedback strategy. We studied how this choice of feedback influenced children's task engagement and robot engagement and their learning gains.

3.5.1 Engagement

The analyses of both engagement types suggest that children seem to be generally engaged with the task and the robot during the three sessions. This accords with human studies indicating that feedback can make tasks encouraging and engaging (Henderlong & Lepper, 2002).

Contrary to our expectations, when the robot provided feedback (either preferred or dispreferred), this did not lead to increased task engagement compared to when the robot provided no feedback (H1a). Children who received no feedback were, on average, rated as equally engaged as children who did receive feedback. However, the type of feedback did seem to have an influence on task engagement of the children: children became more engaged with a robot that provided preferred feedback than with one that used dispreferred feedback or indeed no feedback (H2a). Moreover, the robot's feedback did result into a higher robot engagement compared to no feedback (H1a). Children who received feedback (either preferred or dispreferred), were rated more engaged with the robot than children who did not receive any feedback. However, it is worth pointing out that the numeric effects for task engagement and robot engagement were rather comparable, even though the former but not the latter was found to be statistically significant. Similar to task engagement, children were most engaged with a robot that provided preferred feedback (H2a) in comparison to dispreferred and no feedback. Interestingly, the difference between robot engagement for preferred feedback and dispreferred feedback was larger than the difference for task engagement.

Preferred and dispreferred feedback differed on multiple aspects (variation, extra attempt, repetition of answer, non-verbal behavior) and when combined, these factors seem to have an influence on engagement. While it is hard to identify exactly to what extent each of these factors contribute to children's task engagement and robot engagement, we believe that some aspects might have had a larger effect on both engagement types than others.

For example, variation in feedback, as is realized in the preferred feedback condition, could have had relatively strong effect on children's task engagement and robot engagement. A robot that provides more variation in the way feedback is offered could spark children's interest and keep them interested and motivated in continuing the task over a longer period of time and at the same time also make them more interested in the robot. In contrast, a robot who continually uses the same feedback phrase or no feedback at all might have a negative impact on children's interest in the robot and their robot engagement and moreover reduce their motivation to continue with a task and, thus, be less successful in keeping them task-engaged.

It is furthermore possible that the extra attempt after an incorrect answer in the children's L1 may have task-engaged the children more in the preferred feedback condition than in the other two conditions. The fact that children heard the correct L1 word, could try again and received praise afterwards, may have had a positive effect on their task engagement. This is in line with how teachers tend to provide feedback, praising demotivated children to try to engage them again (Hattie & Timperley, 2007). Some children also mentioned the extra attempt as the robot helping them getting the correct answer, this might increase their sense of relatedness to the robot which could have increased their robot engagement.

Finally, the non-verbal communication of the robot in the preferred condition may have increased children's robot engagement as well. The robot displayed rotating colored eyes and nodded each time when children were correct. This is in agreement with the results of Morris & Zentall (2014), who found that children showed more intrinsic motivation when the robot used non-verbal behaviors such as thumbs up, and the findings of Serholt & Barendregt (2016), who found that children paid most attention to the robot when it provided feedback accompanied by an arm gesture. Future studies that take variation of feedback in combination with different types of non-verbal behavior into account will be needed to develop a full picture of this finding (de Wit et al., 2020). Besides gesturing, also gaze is a known non-verbal factor that can influence engagement (Mwangi et al., 2018). However, in the current experiment gaze was not factor of interest, since the robot's gaze behavior was identical in all three conditions.

As mentioned, it is not possible with the current experiment to determine which factor had the largest effect on task engagement or robot engagement. For this more research is needed. In the current experiment, we explored to what extent by student teachers preferred feedback strategy would differ from a dispreferred feedback strategy or no feedback strategy. We found that preferred feedback has a beneficial effect on both engagement types. However, to identify the effect of different factors that define the preferred feedback strategy has on engagement and which factor contribute to which engagement type, future experiments could be set up in which each factor is varied between conditions.

Also consistent with other studies is that both task engagement and robot engagement seemed to drop over time (de Wit et al., 2018; Kanda et al., 2007; Coninx et al., 2015), and this drop appeared to be similar for all three conditions, although the differences between the conditions stayed over time. Adding more variation to the robot's feedback, as well as varying other parts of its behavior, might help to reduce a drop in engagement. Ahmad et al. (2017) suggested that children seemed to stay engaged with a robot that is adaptive, which lends some support to the importance of individualized variation.

Interestingly, we found an interaction effect between task engagement and the order of feedback strategies but not between robot engagement and order. In particular, we observed that children's task engagement dropped after receiving preferred feedback and that their task engagement was similar or lower before receiving preferred feedback. Receiving no feedback or dispreferred feedback might have demotivated children, and, conversely, receiving various feedback information on their performance, might have increased their motivation again and, therefore, their task engagement. Vice versa, after children received preferred feedback and continued in the dispreferred or no feedback condition, their task engagement decreased again. However, some caution to this explanation must be applied, as the findings might have been influenced by individual differences as well.

3.5.2 Learning gain

As expected, children learned from all three sessions with the robot. They did not learn many words per session though, which is in line with previous research with this young age group (Kory-Westlund & Breazeal, 2015; Vogt et al., 2019). Our results also show that these learning effects were retained in the longer run, because we conducted a post-test two weeks after the last session, suggesting that the target words remained in children's memory (Axelsson et al., 2016).

Contrary to our expectations, children did not learn more in the feedback conditions than when receiving no feedback (H1b), nor did it matter for the learning gain whether feedback was of the preferred or dispreferred variety (H2b). This was not only the case for the post-test, but also applied to the in-game test that was taken immediately after each training round.

What these results suggest is that children could learn from the teaching sessions without the need for feedback, and that the contribution of feedback to learning might have been smaller than we anticipated. This can be explained by the fact that children could rely on cross-situational learning (Smith & Yu, 2008), because children saw four depictions of possible meanings each time they heard a target word, with the distractors changing while the target stayed the same across situations. Hence, children could infer the meaning of a target, even without receiving feedback, based on the co-variation in meanings offered with the different occurrences of the target word, which seems to largely drive the learning, and feedback does not appear to contribute to this learning process.

It is conceivable that the learning task itself might have been too easy for the children to really benefit from the feedback. Moreover, since the children could press any animal they wanted to go forward in the game, they did not have to pay attention to the feedback of the robot. For future research, it would be interesting to conduct a study in such a way that feedback becomes more central to the interaction or more content-related, and where the learning task is more complex (e.g., learning about difficult sentence structures or unfamiliar grammar). This might shed further light on the influence of feedback on learning in child-robot interaction.

It is interesting to note that we did not observe learning differences between preferred and dispreferred feedback, which might be due to the feedback being completely offered in the children's L1. As a result, children did not receive a explicit translation between L1 and L2 as part of their (corrective) negative feedback. This might explain why children did not learn the L2 translation of a concept better during negative preferred feedback. It seems plausible that the addition of L2 to the negative (corrective) feedback would have resulted in higher learning gains (Hall, 2002; Scott & de la Fuente, 2008). However, we did not add this L2-L1 translation to our negative feedback for methodological reasons to keep the different conditions comparable. In particular, we made sure that there was an identical number of L2 exposures in every condition, since the number of L2 exposures could also affect learning (Ellis, 2002).

3.5.3 Relation between engagement and learning

Various studies have found that increased engagement leads to better learning performance (Christenson et al., 2012). However, in our data we did not observe a relation between task engagement or robot engagement and learning. Children who were more engaged with the task or with the robot did not learn more words than children who were less engaged. This might be due to the relatively small learning gain of children in the different conditions. They learned on average close to two out of six words during each session and this might not have been enough to observe a correlation with both engagement types. Moreover, it is conceivable that individual differences between children might have played a role as well. Effects of engagement on learning seemed to differ substantially from one child to the next, which is consistent with earlier research with this age group interacting with a robot (van den Berghe et al., 2021). Finally, we conjecture that in future research with more varied and more prominent feedback (along the lines sketched above), we might indeed observe that more engagement leads to better learning results.

3.5.4 Strengths and limitations

This study has at least four strengths: First, we systematically compared different feedback strategies, derived from actual strategies suggested by young student teachers. Second, we tested a large group of young children to measure the effects of feedback. Third, the study was a carefully constructed experiment, of which all hypotheses and analyses have been preregistered (Simmons et al., 2011). Fourth, we measured two types of engagement to account for the children's engagement with the task and with their engagement with the robot as social partner.

Our study has also at least four limitations. First, we only measured comprehension and not active production of words. However, as speech recognition of the robot is not reliable yet, a more interactive task would have to rely fully on the experimenter in a Wizard of Oz setting (Kennedy et al., 2016). Since we aimed for an autonomously operating system, our task was designed to teach only passive understanding of L2 by using a tablet to record children's responses.

Second, our task was very repetitive. The only variation we introduced was the feedback that the robot would provide in the preferred feedback condition. Children did not have control over when to play with the robot and they were not able to change the task. It is a challenge to design a task that is adaptive to children's preferences, while still being educationally responsible and technical feasible. Providing such autonomy to children could increase their intrinsic motivation, which would increase their engagement and their learning performance (Deci & Ryan, 2000; van Minkelen et al., 2020).

Third, the robot could not react to the children's perceived engagement level during the experiment. While a human teacher would constantly monitor children's engagement and adapt the task accordingly to make it more personalized, the robot in our experiment simply continued to the next word and kept the interaction the same throughout all sessions, disregarding the child's engagement. Being able to automatically recognize a child's engagement would allow the robot to personalize feedback and other behaviors based on this engagement (Gordon et al., 2016; Ahmad et al., 2019).

Finally, we investigated the main effect of feedback on engagement and learning gain and showed that the preferred feedback had an influence on engagement with the task and with the robot. However, preferred and dispreferred feedback varied on multiple factors (variation, extra attempt, repetition of answer, non-verbal behavior), and consequently we cannot attribute the effect on engagement to only one of these factors, only the combination. Future research should look at individual aspects of feedback if technically feasible to measure the effectiveness for engagement.

3.6 Conclusion

The study presented in this chapter explored whether robot feedback affects children's task- and robot engagement and learning gain in second language learning. We compared three robot behaviors: one based its feedback on student teachers' preferred feedback strategies, one that did the opposite and one that did not use any feedback. The preferred strategy varied its feedback, gave children an additional attempt when they answered incorrectly, repeated the target word and gave non-verbal feedback. In contrast, the dispreferred feedback strategy did not vary its feedback, did not provide children with an additional attempt, did not repeat the target word and did not give non-verbal feedback. We found that children in the preferred feedback condition were more engaged than children in the dispreferred feedback and no feedback conditions, both with the task as with the robot. However, the feedback strategy did not influence children's learning gain; they did not retain more word knowledge with one of the different conditions. Moreover, we did not observe a relation between learning and engagement.

Our results are especially interesting for long-term interactions where engagement of children often drops. Providing feedback in an even more varied and motivating manner might help children to remain engaged in long-term scenarios. We expect that in the long-term such varied and motivating feedback can also improve children's learning gains, especially when the learning tasks become more difficult and children cannot just learn from inferring associations through cross-situational learning.

4

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

Based on: de Haas, M., Vogt, P., van den Berghe, R., Leseman, P., Oudgenoeg-Paz, O., Willemsen, B., de Wit, J., & Krahmer, E. (2021) Engagement in Longitudinal Child-Robot Language Learning Interactions: Disentangling Robot and Task Engagement. *International Journal of Child-Computer Interaction*, under revision.

Abstract This chapter investigated a seven sessions interaction between a tutor robot and Dutch preschoolers (5 years old) during which 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, consistent with the novelty effect. In addition, we found that children were less task-engaged with a robot using iconic gestures than with a robot using no iconic gestures when aggregating task engagement over the first sessions (sessions 1-3) compared to the next sessions (sessions 4-6). Interestingly, robot engagement showed the opposite pattern. 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.

4.1 Introduction

Engagement is important for learning (Zaga et al., 2015; Christenson et al., 2012). 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) (Kanero et al., 2018; van den Berghe et al., 2019). 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 et al., 2004; Leite, Martinho, & Paiva, 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 et al., 2017; Kanda et al., 2007; Komatsubara et al., 2014; Leite, Martinho, & Paiva, 2013).

It is important to bear in mind that most long-term HRI studies that studied engagement, only investigated engagement 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 et al., 2014)

It is still unclear whether both task engagement and robot engagement have a positive or a negative effect on 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, therefore, pays less attention to the learning task in front of them (Kennedy et al., 2015). Instead, children interacting with a less distracting robot might pay more attention to the task and become more task-engaged instead of robot-engaged. Thus, task and robot engagement might also be related. 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'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. It is also 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 chapter, therefore aims to add to the evidence regarding the effectiveness of robot-assisted second language learning of pre-school children while specifically focusing on the role of engagement with a social robot, and engagement with the task and the possible positive effect on children's second language (L2) word knowledge.

4.2 Background

4.2.1 Robots in education

Social robots have been used in education for quite some time now (for a review, see Belpaeme, Kennedy, et al., 2018). Most of the times, these social robots are used as a peer tutor, that teaches the child a new skill and can also provide feedback. They have been used with children in many fields, such as, teaching children mathematics or helping children with writing (Hindriks & Liebens, 2019; Konijn & Hoorn, 2020; Kennedy et al., 2015; Alves-Oliveira et al., 2019).

Most interactions between robots and children rely on other devices to get an autonomous interaction since speech recognition has shown to still be unreliable (Kennedy et al., 2017; Mubin et al., 2012). Some child-robot interactions rely on the improvement of speech recognition in the coming years and therefore use a Wizard of Oz approach for their studies (Kory-Westlund, Jeong, et al., 2017). One study used speech recognition and speech-repair-mechanisms such as pressing the buttons on the robot's feet to provide the robot with an answer but found that technical communication breakdowns negatively impacted the interaction (Ligthart et al., 2019). Other researchers used 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, the project Cowriter (Jacq et al., 2016) used a screen for the robot to write on in their child-robot interaction, to teach children how to write with a learning by teaching paradigm. Moreover, Alves-Oliveira et al. (2019) used a screen to display an interactive city map. The robot could interact with the environment, as could the participating children. Ahmad et al. (2017) (2019) used a tablet screen to display a game that children played with the robot. This tablet game was based on an existing board game (Snakes and Ladders), which was implemented in the tablet to increase the autonomous behavior by the robot.

However, these studies did not investigate the difference between an interaction with only a tablet, and an interaction with a tablet and a robot. It is interesting to examine whether the presence of the robot is not distracting from the task, and whether engagement with the robot assists in learning. 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 (Kory-Westlund et al., 2015; van den Berghe et al., 2019). 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 et al., 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. This chapter 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.

4.2.2 Engagement

Despite its common usage, there are multiple definitions of engagement used for HRI. The definition by Sidner et al. (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. Fredricks et al. (2004) argued that engagement is more than only this cognitive aspect. They describe that engagement is built up from three different dimensions: the cognitive, affective and behavioral dimension. This definition shows how complex engagement is and that it measures multiple aspects.

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 (Xu et al., 2016; Chaspari et al., 2015; Chung, 2019), 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 et al., 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 (as discussed in Chapter 2). Sometimes children need to look at the task in front of them, while being engaged, only spend less time looking at the robot. 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 (Jeong et al., 2018; Javed et al., 2020) or in combination with touch behavior on a tablet (Vázquez et al., 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 (Zaga et al., 2015; Díaz et al., 2011), sometimes combined with other techniques such as the use of distractors (Ligthart et al., 2020) or physiological measurements such as thermal infrared imaging (Filippini et al., 2020), or electrodermal activity (Leite, Henriques, et al., 2013) and EEG (Szafir & Mutlu, 2012; Alimardani & Hiraki, 2020; Perugia et al., 2020). 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 (Rich et al., 2010; Ishii & Nakano, 2010; Rudovic et al., 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 (Zaga et al., 2015; Oertel et al., 2020).

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 affect 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 generalise 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 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 feature 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.

ZIKO

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., Storli & Sandseter, 2019; Laevers, 2015; Arnott et al., 2016) 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 (Trowler, 2010): children's levels of concentration, motivation (cognitive dimension), energy (affective), their exploratory drive and persistence (behavioral) 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.

4.2.3 Long-term interactions

Long-term interactions are important to investigate because they look beyond the novelty effect (Leite, Martinho, & Paiva, 2013; Kanda et al., 2007; Ahmad et al., 2019; Oertel et al., 2020). Salter et al. (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 when it used repeated behavior. This can also be confirmed by Serholt & Barendregt (2016) who found that children's social responses to the robot were drastically reduced by the third session. Two other studies investigated primary school children (8-9 years) over time (Leite, Martinho, & Paiva, 2013; Ahmad et al., 2019) and found that their engagement remained the same over time when playing chess five times during five weeks (Leite, Martinho, & Paiva, 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). These studies focused on older children at primary schools and children at that age undergo major developmental changes, which results into 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 Tanaka et al. (2007), Kanda et al. (2007) and the study in Chapter 3. Kanda et al. 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. 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 Chapter three, we found that a robot providing teacher-like feedback had a positive influence on children's task 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.

4.2.4 This study

The current study was part of a large-scale study in which we investigated the effectiveness of a peer-tutor robot 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[1] and 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 tablet and 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 chapter, we present the first longitudinal comparison of robot and task engagement, 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 chapter, we only included the three experimental conditions because the control condition interaction was very different from the other three conditions. We expected that children were more task-engaged when interacting with a robot and a tablet compared to a tablet only (H1). Children to be more more robot-engaged with a robot using iconic gestures than one without iconic gestures (H2). Finally, we expected that children's task engagement positively relates with children's L2 word knowledge (H3a) and children's robot engagement is negatively related with children's L2

[1] https://aspredicted.org/6k93k.pdf

word knowledge (H3b) based on the results by Kennedy et al. (2015).

4.3 Method

4.3.1 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

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 project in which the study was embedded, the L2TOR project, received ethical approval from Utrecht University's Ethics Committee under protocol number FETC16-039.

4.3.2 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 Figure 4.1). It was a between subject design where children received the 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 or 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

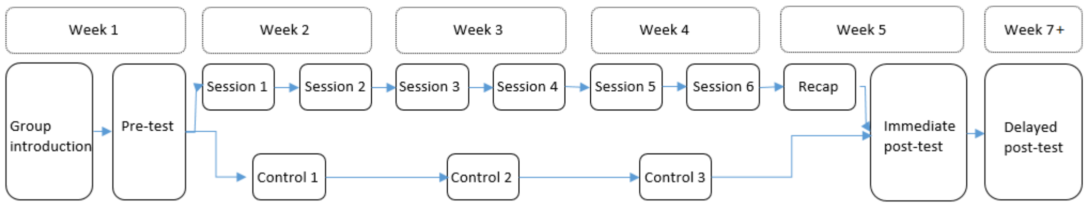


Figure 4.1: Schematic overview of the experiment.

Figure 4.2: Experimental setting

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 chapter, because this interaction was very different than the other interactions and the specific interest of this chapter is children's engagement and the relation with learning and the children in the control condition did not receive the learning activities.

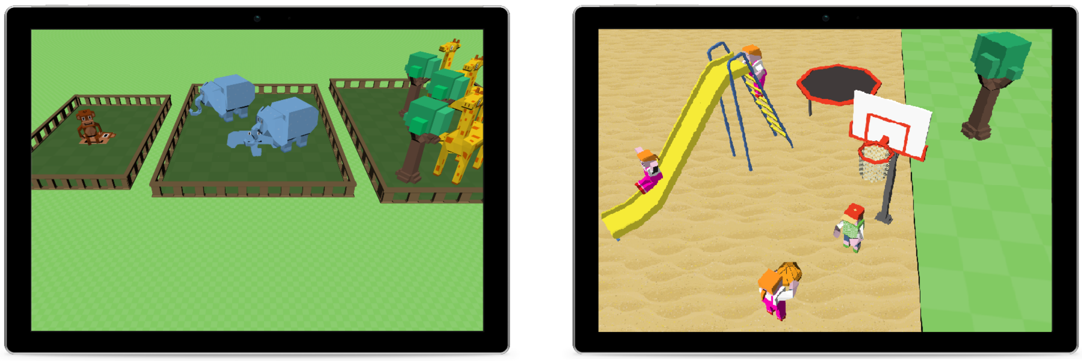
4.3.3 L2 tutoring sessions

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 Figure 4.2). 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 4.1, which were divided in two domains: number domain and spatial domain. The number domain consisted of 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 Figure 4.3a), 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 Figure 4.3b), the tablet displayed a child sliding and climbing a slide.

Table 4.1: 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

Each of the tutoring sessions followed a similar script, and contained a few personalised 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 asked 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,



a: Session 1　　　　　　b: Session 6

Figure 4.3: Tablet environment for session 1 and session 6.

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 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. 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). The interaction was a one-on-one interaction, but the experimenter stayed in the same room to intervene when necessary.

Figure 4.4 shows the number of prompts that children had during each session, children had the least fixed prompts in session 4 and the most in session 7, the recap session. Prompts with the tablet contain actions like dragging and touching objects on the screen, prompts with the robot contain repetition and re-enactment of the target words. The duration of each session was between 15 to 25 minutes. The tablet and robot both instructed the child to execute tasks. After successfully completing a task that was prompted by either the tablet

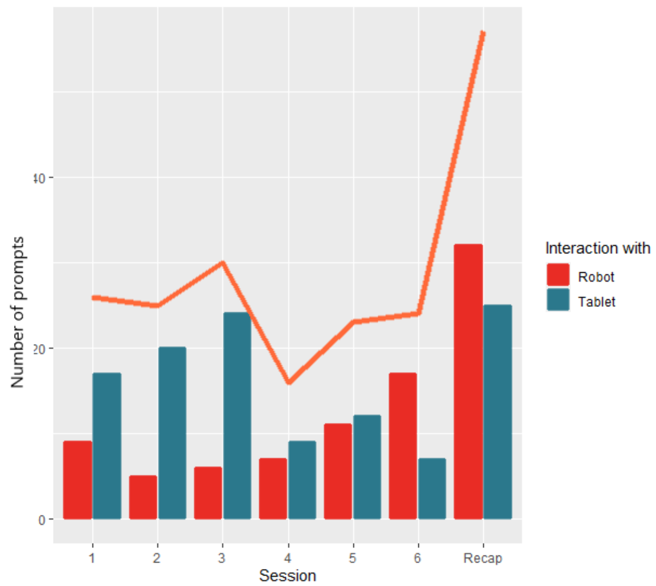


Figure 4.4: Number of prompts per session. The red line shows the total amount of prompts per session. Tablet prompts contain actions like dragging and touching objects on the screen, robot prompts contain repetition and re-enactment of the target words.

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.

4.3.4 Materials

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 non-word repetition skills. Children were asked to translate each target word (34) from 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

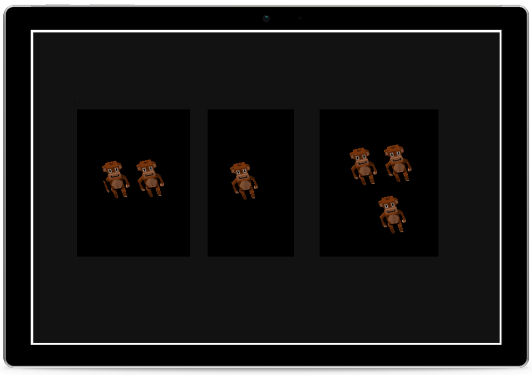


Figure 4.5: In this example, in order to test the children's understanding of the word "two", the tablet asked the children to select the picture showing the two monkeys.

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 et al., 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 et al., 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 chapter. More information can be found about these measurements in Chapter 5.

The total duration of the pre-test was 30-40 minutes. 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 Figure 4.3C) 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 final session and a delayed post-test at least two weeks after the final 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, $\alpha = .87$.

4.3.5 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.

4.3.6 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., to drag 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.

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 minutes per child to get a reliable engagement measure ($r = .83$, Laevers, 2003)($r = .89$, Colpin et al., 2002) for a full day interaction. Because our interactions only took 15 minutes 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 4.3.6 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 4.3.6). It contains example behavior that belonged to a certain engagement level. The scheme is organized in such a way that children who do not show any interest, or are continuously talking to the experiment leader are 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 4.3.6 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[2].

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 into 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).

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

Table 4.2: A part of the engagement coding scheme as used in this experiment. The full coding scheme can be found on Github[2] 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 by 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.

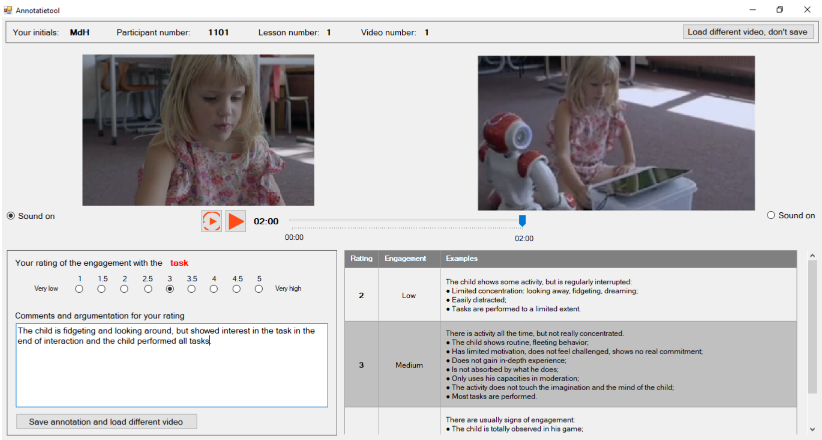


Figure 4.6: Annotation tool for engagement raters.

Annotation software

We developed our own annotation software program for the raters (see Figure 4.6). 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.

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 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 = 0.72, 95% CI[0.70, 0.74])(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 et al. (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. (Note that also in other chapters we have reported similar range scores.). Therefore, we consider the score for this study sufficient for further analysis. We used the raters' weighted average during our analyses.

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 analyses 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 session was already started at the beginning of the video, 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.

4.3.7 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. Therefore, to explore whether children scored higher at the beginning or at the end of the long-term interaction, we combined the first three sessions (1-3) and the following three sessions (4-6) for the post-hoc analysis. We did not include session 7 in this analysis because this was not a word learning session like the other sessions, but designed as a recap session.

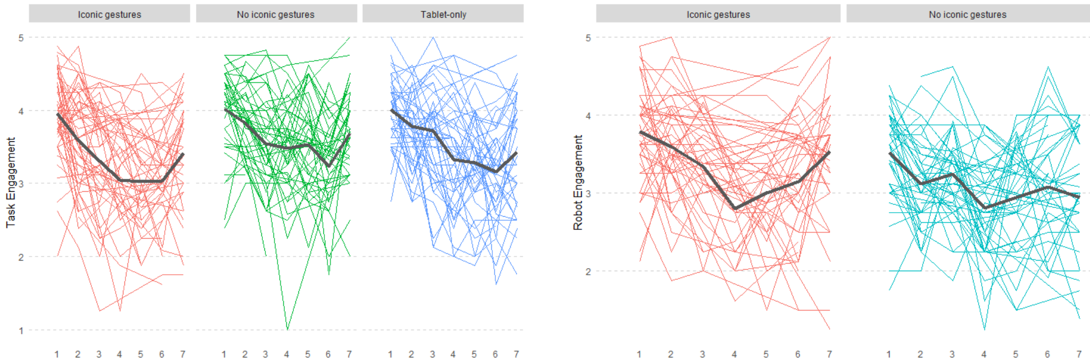
We also explored effects of gender and age on both engagement types, using a t-test to compare the different genders and a linear regression analysis for age. Second, we used Pearson's correlations to investigate how task engagement, robot engagement were related with children's knowledge of L2 words. 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.

4.4 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 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.

Table 4.3: Task and robot engagement scores for each session (SD).

Engagement	Total	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7
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

Figure 4.7: The individual children's engagement ratings over time and per condition. The black line shows the average engagement during each session.

4.4.1 Engagement over time and conditions

Task engagement

Figure 4.7a shows children's task engagement over time for the three conditions. Each line 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.

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 ($F(12, 138) = 0.80, p = .65, \eta^2 = .04$).

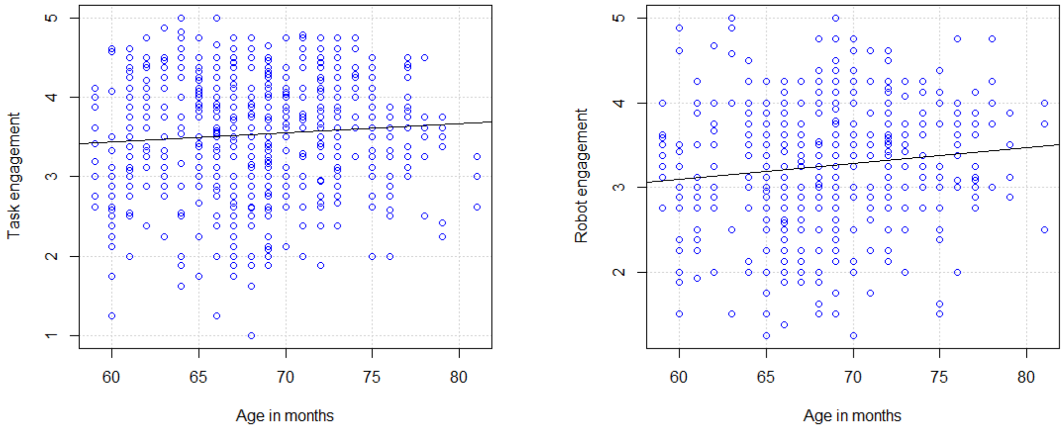
We carried out an exploratory analysis to compare children's task engagement in the beginning of the tutoring sessions with their task engagement in the final sessions. We decided to compare the first three sessions (session 1, session 2, session 3) combined with the following three sessions (session 4, session 5, session 6) combined because it has been reported that children seem to lose their engagement from the third session onward (Salter et al., 2004). We excluded session 7 because this recap session was structured differently from the other sessions which seems to increase children's engagement (cf. Figure 4.7a). A mixed design ANOVA with the averaged task engagement scores as dependent variable, session as within factor (two levels: sessions 1-3 and sessions 4-6) and the condition as between factor showed that children were more task-engaged during the first sessions (1-3) ($M = 3.78, SD = 0.48$) than during the following sessions (4-6) ($M = 3.25, SD = 0.60; F(1, 129) = 122.17, p = .001, \eta^2 = .20$). Additionally, there was a significant effect of condition on task engagement ($F(2, 129) = 4.73, p = .01, \eta^2 = 0.05$). Pairwise comparisons with a Bonferroni correction showed that children were more task-engaged when interacting with a robot without using iconic gestures ($M = 3.66, SD = 0.55$) than when using iconic gestures ($M = 3.39, SD = 0.66, p = .002$) but not more than with only a tablet ($M = 3.57, SD = 0.56; p = .28$). There were no differences between children interacting with a robot using iconic gestures and children interacting only with a tablet ($p = .11$). No significant interaction effect was found between the first and next sessions and condition ($F(2, 129) = 1.09, p = .34, \eta^2 < .01$).

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). Figure 4.8a shows that a younger age was associated with a lower task engagement, however the regression is exceptionally weak.

Robot engagement

Figure 4.7b 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 ($F(1, 13) = 6.74, p = .02, \eta^2 = .15$). 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 ($F(6, 78) = 1.50, p =$



a: Task engagement b: Robot engagement

Figure 4.8: Age plotted against a) task engagement and b) robot engagement.

$.19, \eta^2 = .07$), Moreover, there was no significant interaction effect of robot engagement over sessions in the different conditions ($F(6, 78) = 1.39, p = .23, \eta^2 = .07$).

Additional analysis combining the first three sessions (session 1-3) and comparing those with the combined next three sessions (session 4-6) showed that, similarly as task engagement, there was a significant difference over sessions on robot engagement ($F(1, 92) = 39.30, p < .001, \eta^2 = 0.12$). We found that children were more robot-engaged during the first sessions (1-3) ($M = 3.50, SD = .51$) than the following sessions (4-6) ($M = 3.08, SD = .68$).

Again, there was a significant difference between the two conditions for robot engagement when comparing the three first sessions with the last sessions ($F(1, 92) = 6.58, p = .001, \eta^2 = .05$). Children were more robot-engaged with a robot using iconic gestures ($M = 3.43, SD = 0.68$) than with a robot without iconic gestures ($M = 3.16, SD = 0.54$). However, there was no significant interaction between sessions and conditions on robot engagement ($F(1, 92) = 0.02, p = .90, \eta^2 < .01$).

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 ($t(527) = 0.86, p = .39$).

Moreover, 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). Figure 4.8b shows that a younger age was associated with a lower robot engagement, however the explained variance is very small.

Table 4.4: An overview of children's word knowledge scores. 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.

4.4.2 Relation engagement and word knowledge

Table 4.4 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 4.5 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 atten-

Table 4.5: 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: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

tion 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.

Table 4.6: Correlations between active interaction moment during the game and children's engagement.

Interactions with	Task engagement	Robot engagement
Tablet	.21***	.18***
Robot	-.11**	-.04
Total	.04	.07

Note: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

4.4.3 Relation engagement and fixed interaction during sessions

We also explored the relation between the fixed game play 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) according to Figure 4.4. As Table 4.6 shows, children's task and robot engagement both correlated positively with the fixed interaction moments with the tablet interactions and negatively with the fixed interaction with the robot. Both children's task and robot engagement significantly correlated when children had to interact more often with the tablet (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 ($r(807) = -.11, p = .002$). Note that the required game play only refers to the fixed interaction moments between the game and child (manipulating objects on the tablet, required verbal and non-verbal behavior toward the robot), and not to the unscheduled or passive interaction between the game and child (e.g., the robot's feedback, or the tablet displaying content).

4.5 Discussion

The aim of the present chapter 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, (b) a robot without iconic gestures, 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 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, yet supported when comparing sessions 1-3 with sessions 4-6 that showed a significant decrease. Both engagements seemed to fluctuate less after the third session, which 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.

4.5.1 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; Serholt & Barendregt, 2016; Leite et al., 2014).

Contrary to our expectation (H1), children were not more task-engaged in the robot conditions than in the tablet-only condition. We explored the effect of the conditions when the novelty effect plays of less a role by comparing the first three sessions combined with the next three sessions. When we compared the mean of the first sessions with the last sessions, we did find an effect of condition on children's task engagement. Children were more task-engaged when interacting with a robot using no iconic gestures than with a robot using iconic gestures, although the effect size was very small. It seems possible that this result is due to that the robot using iconic gestures attracted attention away from the task and therefore reduced children's task engagement (in line with Kennedy et al., 2015). Strangely this effect is only be seen when we compare the first sessions (1-3) with the following sessions (4-6) to reduce the influence of the novelty effect.

It is not entirely clear why this effect was not observed when taking all sessions apart. One possible explanation is when we examined the data in further detail, it showed that children's task engagement decreased more rapidly during the first sessions when playing with a robot using iconic gestures than without iconic gestures. This observed difference may be due to the nature of the iconic gestures. The gestures occurred very frequently and this increased the duration of each session considerably, which may have negatively affected the task engagement when children were interacting with a robot using iconic gestures compared to a robot without iconic gestures. Another possible explanation is that there were large individual differences between children in task engagement over sessions, something that we also found in the other chapters. 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 et al. (2021) discussed the individual differences of this study in more detail.

We did not include the recap session in this analyses because this session was very different than the other sessions and children's task engagement increased during this session. 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.

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 fixed interaction moments in the sessions, children's task engagement was, as expected, positively related to the tablet's fixed interaction moments (e.g., dragging an object on the screen, selecting an object on the screen). Interestingly, there was a negative relation between children's fixed interaction with the robot (speech and re-enactment of gestures) and children's task engagement. This was unexpected because these fixed interactions with the robot were also part of the task. This negative relationship may possibly be explained by that these active 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 et al., 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.

4.5.2 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 H2). 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 toward the robot.

Unlike task engagement, children's robot engagement did not decrease significantly over the different sessions. Only when comparing the first three sessions (1-3) with the following three sessions (4-6), their overall robot engagement dropped. 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 fixed interaction moments. Session four had the fewest interaction moments 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 non-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 fixed tablet interaction moments but surprisingly, no significant correlation between robot engagement and the robot's fixed interaction moments. It is difficult to explain these findings, but it is important to note that these interaction moments only focused on the fixed interaction moments (i.e. when the children had to respond in one way or another) as implemented in the game. This positive correlation between robot engagement and the tablet fixed interaction moments may actually be due to the robot's feedback. During the tablet interaction moments, 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 fixed robot interactions, but always followed the fixed tablet interactions. Arguably, this positive feedback by the robot increased children's robot engagement. This also accords with observations in Chapter 3, 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 required interaction moments. 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.

4.5.3 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 et al., 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. Our findings show that there is a 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.

4.5.4 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 generalise 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.

Our study also has several strengths. It is one of the first studies to investigate task 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 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.

4.6 Conclusion

In this chapter, we present one of the first large-scale studies that investigated children's task 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 do have an effect on children's task and robot engagement over time. Task engagement was higher with a robot that did not use iconic gestures, compared to one that did. Robot engagement showed the opposite pattern: Children were more robot-engaged when a robot used iconic gestures than without iconic gestures. Moreover, children's task 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 and robot engagement in long-term child-robot interactions.

5

Anthropomorphism and the relation with second-language learning in longitudinal child-robot interaction

Based on: van den Berghe, R.*, de Haas, M.*, Oudgenoeg-Paz, O., Krahmer, E., Verhagen, J., Vogt, P., ... & Leseman, P. (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. https://doi.org/10.1111/jcal.12497
**Rianne van den Berghe and Mirjam de Haas had equal contributions*

Abstract This study investigates the degree to which children anthropomorphize a robot tutor and whether this anthropomorphism relates to their vocabulary learning in a second-language (L2) tutoring intervention. With this aim, an anthropomorphism questionnaire was administered to 5-year-old children ($N = 104$) twice: prior to and following a seven-session L2 vocabulary training with a humanoid robot. On average, children tended to anthropomorphize the robot prior to and after the sessions to a similar degree, but many children changed their attributed anthropomorphic features. Boys anthropomorphized the robot less after the sessions than girls. Moreover, there was a weak but significant positive correlation between anthropomorphism as measured before the sessions and scores on a word-knowledge post-test administered the day after the last session. There was also a weak but significant positive correlation between the change in anthropomorphism over time and scores on a word-knowledge post-test administered approximately 2 weeks after the last session. Our results underscore the need to manage children's expectations in robot-assisted education. Also, future research could explore adaptations to individual children's expectations in child-robot interactions.

5.1 Introduction

5.1.1 Anthropomorphism

When interacting with a social robot, people have a tendency to attribute human forms, characteristics, and/or behaviors to the robot. This phenomenon is called anthropomorphism (Bartneck et al., 2009). People do not only anthropomorphize robots, but also many other non-human entities, such as animals, toys, and machines (Caporael, 1986), and presumably this helps them to understand and gain control over their environment (Duffy, 2003; Waytz et al., 2010). Anthropomorphism can be a useful mechanism in human-robot interaction (Duffy, 2003; Fink, 2012), because people evaluate robots more positively, collaborate better with them, and empathize more with robots that are more human-like or display more human-like behavior than with robots that are less human-like (Breazeal et al., 2005; Eyssel et al., 2012; Hegel et al., 2008; Moon et al., 2014; Riek et al., 2009). In this chapter we set out to study the degree to which children anthropomorphize a humanoid robot, how children's anthropomorphic beliefs about the robot may change after multiple interactions with the robot, and whether children's anthropomorphistic perception of the robot and word knowledge after a second language (L2) vocabulary training are related.

The degree to which people anthropomorphize a robot is affected by the robot's appearance and behaviors (DiSalvo et al., 2002; Phillips et al., 2018; Tung, 2016). For example, people are more likely to anthropomorphize robots that have a torso, a skin, or appear to have gender (Phillips et al., 2018). Robot movement in general has also been found to increase human-likeness ratings (Tung, 2016). More specifically, using co-speech gestures has been found to increase anthropomorphism, and the use of social gaze to increase life-likeness (Salem et al., 2013; Zaga et al., 2017). However, people do not all anthropomorphize robots to the same degree. One of the reasons for these individual differences is that people use their own experiences in rationalizing the actions of an object and in reasoning about its mental states (Epley et al., 2008, 2007; Lemaignan, Fink, & Dillenbourg, 2014), and may thus ascribe different mental states to objects depending on their own experiences. Thus, in human-robot interaction, the degree to which people anthropomorphize robots likely does not only depend on the type of robot used and the behavior the robot displays, but also on the specific characteristics and experiences of the person interacting with the robot. While most robot research on anthropomorphism has focused on adults (see Fink (2012) for a review), children of all ages have been found to anthropomorphize robots as well (Beran et al., 2011; Kahn Jr et al., 2013; Lemaignan et al., 2015; Monaco et al., 2018). Younger children (up to twelve years old) are more likely than older children to anthropomorphize robots (Beran et al., 2011; Kahn Jr et al., 2012; van Straten, Peter, Kühne, & Barco, 2020). They experience more enjoyment and are less sensitive to the robot's style of interaction than older children (van Straten, Peter, & Kühne, 2020), which may relate to a higher degree of anthropomorphism. In particular, younger children are more likely to assign cognitive and affective beliefs to robots than older children, such as the ability to remember people and understand people's feelings (Beran et al., 2011). However, even preschool children attribute few biological properties to robots (Jipson & Gelman, 2007) and already understand that robots are something in between living beings and mechanical artifacts (Kory-Westlund & Breazeal, 2019a). In a meta-analysis by van Straten, Peter, & Kühne (2020) a robot's responsiveness and role were the strongest predictor of children's closeness to a robot but the predictors for trust were not consistent. Also, this meta-analysis showed that boys feel more close to a robot with the same gender but girls are not affected by the gender of the robot.

5.1.2 Changes in anthropomorphism

Previous research indicates that children's perceptions or expectations of robots can change over time. Children value a robot's properties differently depending on their experience with robots (Obaid et al., 2015; Sciutti et al., 2014). Before interacting with a robot, children attribute more importance to a robot's shape (e.g., having a head or arms) than its sensory and motor properties (e.g., the ability to feel or move). After having interacted with a robot, they value its sensory and motor properties more and its shape less than before (Sciutti et al., 2014). While Sciutti et al. (2014) did not specifically investigate anthropomorphism, it does suggest that sensory and motor properties, which can be linked to anthropomorphism, may become more important over time when children's experience with robots increases. Bernstein & Crowley (2008) asked children between four and seven to evaluate different entities (including two robots) on livingness and intelligence. Children who had had little experience with robots, judged the robot more often as living than children who had had more experience with robots. Moreover, children who had had experience with robots were more likely to distinguish robots from other entities that they already knew (e.g., things that are living) and consider robots as intelligent, albeit in a unique "robot intelligence" manner. In contrast, a study by Kory-Westlund et al. (2016) did not find changes in anthropomorphism. A robot was framed either as a social agent or a machine by using either inclusive language and second-person pronouns or third-person pronouns and the word "robot". In this study, children between ages three and seven played a sorting game with the robot. The degree to which they anthropomorphised the robot was assessed through a questionnaire administered both before and after the game. The study did not show an effect of framing on children's anthropomorphism, and there was no difference in the degree to which children anthropomorphised the robot before or after the game. It is not clear from these studies whether children's anthropomorphism is indeed unaffected by their interaction with the robot, or whether one interaction session was not enough to change their degree of anthropomorphism. On the one hand, people might attribute cognitive and social abilities to robots that they cannot meet (Dautenhahn, 2004), which is particularly a problem for repeated interactions (Leite, Martinho, & Paiva, 2013). On this idea, the longer people would interact with robots, the more likely it should be that the robot falls short of these expectations, which would negatively affect people's tendency to anthropomorphize the robot. Evidence for this idea comes from a previous study with children in which explicitly informing children on the robot's lack of psychological abilities (e.g., self-consciousness, social cognition) led to lower anthropomorphism and trust (van Straten, Peter, Kühne, & Barco, 2020). It is also in line with a proposed model on the dynamics of anthropomorphism (Lemaignan, Fink, & Dillenbourg, 2014; Lemaignan, Fink, Dillenbourg, & Braboszcz, 2014). In this model, people are most likely to anthropomorphize a robot when first encountering it, because of their expectations about the robot and because the robot's behavior may seem unpredictable and complex. Upon getting acquainted with a robot, people build a mental model to predict the robot's behavior, and as the accuracy of this model increases, the robot is considered more machine-like than human-like, and anthropomorphic tendencies decrease. On the other hand, studies have found that children attributed more anthropomorphic or more positive judgments after having repeated interactions with a robot (Leite et al., 2017; Michaelis & Mutlu, 2018). Michaelis & Mutlu (2018) had ten- to twelve-year-old children participate in in-home guided reading activities with a robot, and found that more children attributed feelings, emotions, and a personality to the robot after the two-week study than before. Though not measuring anthropomorphism directly, Leite et al. (2017) focused on likeability and found that four- to ten-year-old children liked the robot more after having multiple conversations with it. The study in this chapter is aimed at further investigating changes in children's evaluations of a robot in terms of anthropomorphism after multiple interactions with this robot, and relating these evaluations to their learning outcomes in a vocabulary training.

5.1.3 Anthropomorphism and learning

Education is one of the most widely used domains in which social robots are used. Robots can be used to support children's learning, and as such, complement teachers. One of the most often used applications is the use of a robot as a tutor, such that the robot and child together work through educational materials and the robot provides individual support to the child (Belpaeme, Kennedy, et al., 2018). A robot can interact with the children in their physical, referential world. The robot's embodiment and its potential for social interactions to establish common ground is one of the advantages social robots in theory have over other forms of technology such as tablets (Belpaeme, Kennedy, et al., 2018). Physical robots indeed have generally been found to be more enjoyable and a preferred social partner compared to their virtual counterparts (Kidd, 2003; Pereira et al., 2008). It is assumed that such robots are more natural conversational partners, and robot-assisted learning interactions may benefit from similar social behaviors as humans use in learning interactions, such as the use of gestures (de Nooijer et al., 2013; de Wit et al., 2018; Kelly et al., 2009; Macedonia et al., 2011; Tellier, 2008; Verhagen et al., 2019). Furthermore, children have been shown to be less anxious and more motivated when learning with a robot than without a robot (Alemi et al., 2015). Finally, an advantage of a robot is that it can endlessly repeat tasks with individual children where a teacher has to pay attention to other children.

These advantages of robots in education may particularly benefit robot-assisted language learning, which is studied in this chapter. Robots can gesture, move around, and manipulate objects, and by doing so, embed the language that they are teaching in the physical environment that they share with the learner. For example, robots can point to the objects they are naming or act out the meaning of a word. This embedding is known to be important for language learning (Barsalou, 2008; Hockema & Smith, 2009; Iverson, 2010; Oudgenoeg-Paz et al., 2015; Wellsby & Pexman, 2014). As a result, (second) language learning has often been studied robot-assisted learning research (see van den Berghe et al. (2019); Kanero et al. (2018) for reviews). So far, results on the effectiveness of robots for language learning are mixed, however. In this chapter, we further explore one of the factors that may, at least in part, explain the mixed findings in earlier work, but has received relatively little attention to date: anthropomorphism. As discussed earlier, anthropomorphizing robots seems advantageous for human-robot interactions (Duffy, 2003; Fink, 2012), but it is not clear if and how anthropomorphism can affect robot-assisted (language) learning. Yet, the degree to which learners anthropomorphize robots may play an important role in learning situations too, as learning is first and foremost a social process (Vygotsky, 1978). Children who anthropomorphize the robot to a greater degree might interact with the robot in ways similar to how they would interact with peers. Peer learning has been shown to be beneficial to learning (see Topping (2005) for a review), either directly through helping each other, or indirectly through enhancing motivation, confidence and enjoyment. Anthropomorphism is related to social presence: "the degree to which a user feels access to the intelligence, intentions, and sensory impressions of another" (Biocca (1997), Section 7.2). It reflects paying attention to each other, understanding each other, and adapting behavior and emotions towards each other. It is no surprise that such values are also crucial to successful vocabulary training (Marulis & Neuman, 2010), and may thus apply to the robot-assisted vocabulary training in this chapter. It may be worthwhile to design robots in such a way that they make learners feel as if it has a social presence, but the learner's perception of the robot and its social presence may be just as important. It is possible that a robot's benefits as a peer learner or tutor depend on the degree to which the learners anthropomorphize it. In other words, it is possible that a robot perceived as more human-like is more effective when learning a second language than a robot that is perceived as a machine. This begs the question if and how anthropomorphism and learning are related to each other, which is the central research question of this chapter.

Research that comes closest to answering this question is that of Chandra et al. (2018). This study did not directly focus on anthropomorphism, but the researchers did measure children's perception of a robot in terms of intelligence, likeability, and friendliness, and whether this affected their learning in a learning-by-teaching paradigm. In this study, twenty-five seven-to-nine years old children taught a NAO robot to write over the course of four sessions as a way to improve their own writing. There were two conditions: (1) the robot improved its handwriting for half of the children, and (2) the robot did not improve its writing for the other half of the children. Children in the first condition were able to perceive the robot's improvement by the last session, but this as such did not change how they perceived the robot's intelligence, likeability, and friendliness. However, children's own improvement in writing was positively correlated with the likeability of the robot. In the condition in which the robot did not improve, children's perceptions of the robot's intelligence, likeability and friendliness did not change either, but in this condition children's own learning was correlated with the perceived friendliness of the robot. These findings need to be interpreted with caution because of the small sample size and because they did not measure anthropomorphism, but they suggest that children's perception of the robot may indeed be related to their learning. Our study expands on this previous work. It includes an L2 vocabulary training of multiple sessions, thus enabling us to study children's anthropomorphism of a robot and changes therein over a longer period of time. This increases ecological validity, as robot-based interventions aimed at teaching children a particular topic usually span a few weeks, causing novelty effects of the robot that wear off after multiple interactions (e.g., Kanda et al., 2004). We assess the degree to which children anthropomorphize the robot both before and after having interacted intensively with it, allowing to observe changes in anthropomorphism, and examine how children's anthropomorphism and changes therein relate to language-learning performance.

5.1.4 This study

The current study was part of the L2TOR project, which evaluated the effectiveness of a multiple-session L2 learning intervention for young children using a social robot in a large-scale randomized control trial (Vogt et al., 2019) and is the same group of children as in Chapter 4. This long-term control study was pre-registered on AsPredicted[1] and included four conditions: (1) an L2 vocabulary training with a tablet and a robot that performed iconic and deictic gestures to support word learning (gestures that visualize target words and pointing gestures), (2) an L2 vocabulary training with a tablet and a robot without iconic gestures (only pointing gestures), (3) an L2 vocabulary training with a tablet only (no robot involved), and (4) a control condition in which children only played dancing games with the robot. Word knowledge was tested on three occasions, during a pre-test, an immediate post-test, and a delayed post-test (administered between two and four weeks after the training). The results of this pre-registered study regarding children's word knowledge are reported in Vogt et al. (2019) and showed that, irrespective of condition, children knew significantly more words after the tutoring sessions than before. Moreover, children in the experimental conditions (robot with iconic gestures, robot without iconic gestures and tablet-only) scored significantly higher than children in the control condition on word-knowledge tests during the immediate and delayed post-tests. There were no differences between the experimental conditions, such that children who had taken the tutoring sessions with the robot (with or without iconic gestures) did not know more words than children who had taken the sessions with the tablet only. In the current chapter, we only included the experimental robot conditions (i.e., conditions 1 and 2) to investigate the degree to which children anthropomorphized the robot and the way in which this relates to their word knowledge. In our analyses,

[1] https://aspredicted.org/6k93k.pdf

we did not include the tablet-only and control conditions because, children in these conditions either did not interact with the robot (tablet condition) or were not taught any English words by the robot (control condition). We addressed the following research questions and hypotheses:

1. Are there individual differences in the degree to which children anthropomorphize the robot? We expect children to differ in the degree to which they anthropomorphize the robot, in line with previous research on individual differences in anthropomorphism (Epley et al., 2007, 2008).

2. How does the degree to which children anthropomorphize the robot change through multiple L2 tutoring sessions with the robot? Although the evidence is mixed (e.g., (Bernstein & Crowley, 2008; Kory-Westlund et al., 2016; Michaelis & Mutlu, 2018), we expect that anthropomorphism will change over time in different ways, due to the multiple interactions children have with the robot. On the one hand, children may come to perceive the robot more as a friend after repeated interactions, thus perceive the robot as more human-like. On the other hand, it is also possible that children initially have high expectations of the robot's interactive qualities, which the robot, however, cannot meet. In that case, their perception would change over time towards considering the robot as less human-like.

3. How are children's anthropomorphistic perceptions of the robot and their knowledge of L2 words related? We expected word knowledge and attributing human-like cognitive, emotional, and biological qualities to the robot to be positively related to each other. Specifically, we anticipated that children who would anthropomorphize the robot more would treat the robot as a peer that has social presence, and, as such, benefit more from its presence in terms of increased motivation and engagement, that, in turn, would foster word learning. It should be noted that while this design does not enable us to study causal relations between anthropomorphism and word knowledge, we do study whether the two are related and therefore provide evidence pertinent to the possible role anthropomorphism can play in the effectiveness of robot-based educational interventions.

5.2 Method

5.2.1 Participants

This study reports on a part of the sample described in Vogt et al. (2019) and Chapter 4, that is, the children in the two experimental robot conditions. Data was used from 104 monolingual Dutch children (50 girls, 54 boys) with an average age of 5 years and 8 months (*SD* = 5 months) who followed the vocabulary training in one of the two robot-assisted conditions (with or without iconic gestures). These children were recruited from the kindergarten of nine primary schools in the Netherlands. Within schools, children were randomly assigned to one of the conditions, while ensuring a similar gender distribution over the conditions.

1. Robot with iconic gestures: $N = 53$, $M_{age} = 5$ years and 8 months, $SD = 5$ months, 30 boys and 23 girls
2. Robot without iconic gestures: $N = 51$, $M_{age} = 5$ years and 8 months, $SD = 5$ months, 25 boys and 26 girls

Sixteen additional children were excluded when they: (i) knew more than half of the target words in the pre-test (*n* = 3), (ii) did not complete the experiment due to technical issues (*n* = 2), (iii) did not want to participate anymore (*n* = 8), and (iv) did not complete the anthropomorphism questionnaire during the pre-test (*n* = 3). All children's parents signed an informed consent form to allow their children to participate in this study. Children received a small gift at the end of the study to thank them for participation. 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.

5.2.2 L2 tutoring sessions

The aim of the L2 tutoring sessions was to teach each child 34 English words in the domains of mathematical and spatial language. Each child received seven tutoring sessions involving the robot and a tablet. During each of the sessions children were introduced to five or six new target words. The Softbank Robotics NAO robot was used, which was sitting in a 90-degree angle next to the child (see Figure 5.1). A three-dimensional game was developed for the tablet, in which a particular scenario was displayed (e.g., animals in the zoo that had escaped their cages). This served as the context in which the L2 words were introduced (see Figure 5.2). For each word, the child and the robot had to perform different tasks on the

Figure 5.1: A child playing with the robot.

tablet (e.g., selecting or dragging objects on the screen, repeating target words out loud, or acting out target words). For instance, the robot would ask the child to drag three escaped animals back into their cage on the screen. While dragging, the robot would count in English the number of animals in the cage. During these tasks, the robot acted as a slightly more knowledgeable peer who was also being taught English words, but could provide feedback on the child's actions when needed. For example, when a child dragged the wrong animal to a cage on the tablet, the robot could ask the child to drag the correct animal to the cage. See Table 5.1 for an example of the child and robot interaction.

The sessions were designed without relying on children's speech because speech recognition is currently still unreliable with children (Kennedy et al., 2017). For the few times children had to repeat a word, a Wizard of Oz was used where the researcher pressed a button on a control panel after the child had repeated after the robot. The rest of the interaction was carried out autonomously. The interaction was one-on-one in a separate room, but the experimenter stayed in the same room to intervene when necessary and to control the Wizard.

5.2.3 Robot behavior

During the sessions the robot was in breathing-mode (moving with its arms) to appear more lively. As the robot motors can be quite loud when the robot moves, the breathing-mode

Table 5.1: An example of an interaction between robot and child

Robot	Child's action
Tablet shows an environment with three cages, and three giraffes outside one of the cages	
Let's put the <giraffe>in its cage!	drags giraffe in cage
Well done!	
There are still <two giraffes>outside of the cage. There are <more giraffes>outside of the cage than inside the cage. Can you <add one>giraffe?	drags giraffe in cage
Well done!	
We had the <add one giraffe>and now there are <two giraffes>in the cage. There are <more giraffes>inside the cage than outside the cage. Can you add <one giraffe>?	drags giraffe in cage
Well done!	
Please touch the cage with the three animals, so we can hear what three is in English	touches cage with the three giraffes
Tables says <three>	
Repeat me: <three>	
	says three
Well done! <........ >	
Tablet shows adds three trees to the tablet environment	
Cool! The last thing we need to do is put food in the cage with the giraffes. This cage has the <most>animals so they need the <most>trees. Put the trees in the cage so the giraffes can eat from them. We have <three>giraffes, so we need <three>trees. Put the trees in the cage. Count them while dragging	Drags first tree
Let's do one more	Drags second tree
And the last one	Drags third tree
Well done!	
Great! Now each giraffe has their own tree because there are <three>trees and <three>giraffes. The cage is pretty full because <most>animals are in the giraffe cage with the <most>food. You did very well! Let's do something else!	

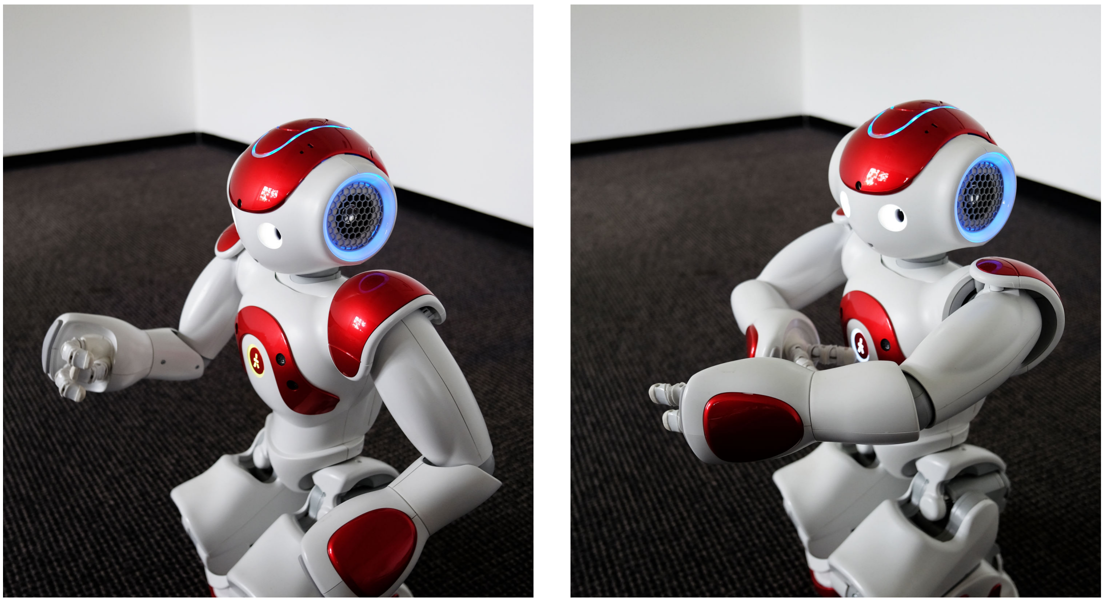
Note: The whole interaction was in Dutch, except for the words between brackets<>.

Figure 5.2: Example of one of the virtual environments that was used as a context for the language-learning interaction

also reduced the initial sound shock when the robot was going to make a gesture and moved up its arms. In both conditions the robot used deictic gestures, such as pointing, to draw the child's attention to the tablet, and head movements to look at the child when the child was asked to perform a certain task on the tablet. The only addition to the iconic + deictic gesture condition on top of the deictic gesture condition was the robot's use of iconic gestures. Specifically, an iconic gesture was designed for each of the included target words, and the robot would perform this gesture whenever it produced that word in the L2. Gestures were designed using key framing (Pot et al., 2009), an animation technique where the designer defines a number of key positions of a character's limbs, and smooth transitions between these points are automatically generated. The design was based on human-performed gestures, which were recorded by means of a gesture elicitation procedure where participants were asked to come up with a gesture depicting each of the target words. The resulting robot gestures were recreated based on the recorded examples, while taking into account the robot's physical limitations (such as its inability to move individual fingers) and the fact that the robot would be sitting down rather than standing, as the human performers were. Figure 5.3 shows examples of the robot gestures for the target words running and behind.

5.2.4 Materials and measurements

Anthropomorphism questionnaire This anthropomorphism questionnaire was constructed for the purposes of the present study and administered by an experimenter in a one-



a: Running   b: Behind

Figure 5.3: Examples of iconic gestures used in this study, photographed from a position where the child would sit. (a) *Running* is gestured by moving both arms back and forth as if the robot is running. (b) The word *behind* is gestured by moving the right hand up and down behind the left hand. Figures taken from (Vogt et al., 2019).

on-one session with the child. The questionnaire took about ten minutes to complete. It consisted of twelve questions (for an overview, see Table 2 in the Results section) and assessed the degree to which children anthropomorphized the robot with regard to various types of properties: biological (e.g., feeling pain, need for food, and ability to grow), cognitive (e.g., thinking, remembering), and emotional (e.g., being happy, being sad). Each question could be answered with "yes"/"no"/"I don't know" and was followed by an open-ended query asking children why they gave this response. The items were based on Jipson & Gelman (2007), who investigated to what extent children make a distinction between living and non-living items. The questionnaire was adapted to fit the present study by adding several items to more thoroughly assess anthropomorphism (e.g., rather than measuring the robot's emotional abilities by only asking whether the robot could feel happy, an item was added on whether the robot could feel sad). The children's closed-ended answers were compared with the open-ended answers to find out whether the children understood the question. Two of the included questions (i.e, "Can the robot break?" and "Is the robot made by humans?") proved unreliable as children's answers to the open-ended query did not correspond to their answers on the close-ended questions. Therefore, we removed these items from our analysis. The children were awarded one point for each "yes"-answer, which indicated that they attributed human-like properties to the robot, and their anthropomorphism score was the proportion of "yes"-answers. We used proportions rather than total scores because there were missing values on some items for some children. This was the case for one child at the pre-test (four of the twelve questions were not administered) and for five children at the post-test (for each of whom one question was not administered). Thus, the maximum score was 1, with a score closer to 1 denoting a child's tendency to consider the robot as human-like. Cronbach's alpha indicated that the internal consistency of the questionnaire was satisfactory, $\alpha = .72$ at the pre-test and $\alpha = .75$ at the post-test.

Comprehension test The comprehension test was a picture-selection task. In this test, children were presented with a prerecorded target word and asked to choose which one out of three pictures or short video clips matched this word best ("Where do you see: [heavy]?"). Each target word was presented three times with different target and distractor stimuli in random order to decrease the chance of children guessing the correct answer. Only half of the 34 target words that were presented in the vocabulary training were included, as a test including all target words would have taken too long for these young children. The same test was used in both post-tests. The internal consistency of the comprehension test was good, with Cronbach's alpha $\alpha = .84$ at the first post-test and $\alpha = .87$ at the second post-test.

Additional measures In addition to the anthropomorphism questionnaire and comprehension task, we administered several tasks assessing general cognitive skills. These tasks are beyond the scope of this study as they did not assess anthropomorphism (see Vogt et al. (2019)). They were: (1) a Dutch receptive vocabulary test (Peabody Picture Vocabulary Test; Dunn et al. (2005)), (2) a selective attention task (visual search task; Mulder et al. (2014)), and (3) a phonological memory test (quasi-universal nonword repetition test, Boerma et al. (2015)). Moreover, we administered two translation tests to measure children's knowledge of the English words, in which children listened to the target words in L2 and were asked for their Dutch translations, or vice versa. The English-to-Dutch translation test was used as a pre-test. Note that the main purpose of this translation test during the pre-test was to enable us to exclude any children who knew many words prior to the lesson series, although it also allowed us to compare pre- and post-test scores (see Vogt et al. (2019) for these analyses). We chose not to include a comprehension test as a pre-test, as children may learn from such tests, given that, unlike in the translation task where no answer is provided, a word is presented with pictures, one of which depicts the word's meaning. Moreover, in this chapter, we did not include the translation tests in the analyses, as there was low variability in children's scores. Thus, in this chapter, we only include the comprehension test as a measure of children's word knowledge.

5.2.5 Procedure

Prior to the experiment, all children participated in a group introduction with the robot to familiarize the children with the robot, build trust, and explain the basic similarities and dissimilarities between the robot and humans (e.g., the robot speaks without moving its mouth, but looks at us while speaking in the same way as humans do; Belpaeme, Vogt, et al. (2018)). These explanations were deemed necessary to make sure that children would know how to interact with the robot in the subsequent sessions. During the introduction, participants danced together with the robot, were allowed to shake the robot's hand, and played a brief gesture imitation game. The robot was not explicitly framed as either a human or a machine, by avoiding pronouns and by being called "Robin the robot" (i.e., a combination of a gender-neutral human name and the label "robot"). After the introduction, a pretest was administered including the anthropomorphism questionnaire and several tests measuring general cognitive skills as well as children's knowledge of the English words. In the weeks thereafter, the children received seven one-on-one tutoring sessions with the robot. Each session took approximately 17 minutes to complete. One or two days after the last session, an immediate post-test was administered including the anthropomorphism questionnaire for the second time, the comprehension test, and other tasks measuring children's knowledge of the English words. Finally, a delayed post-test was administered in which the comprehension test and other English vocabulary tasks were repeated, between two and four weeks after the tutoring sessions.

5.2.6 Analyses

In the results section, each research question is addressed in a separate paragraph. First, we examined whether there are individual differences in the degree to which children anthropomorphize the robot before the tutoring sessions (RQ1). We used independent-samples t-tests to explore effects of gender and condition, and a linear regression analysis for age. We used age as a continuous variable in our analyses, but reported means for a "younger" and an "older" age group in Table 5.3 in the results section, calculated through a median split (at 68.2 months). Second, we investigated how the degree to which children anthropomorphized the robot changed through multiple L2 tutoring sessions with the robot (RQ2). We used a paired-samples t-test to compare anthropomorphism scores before and after the tutoring sessions. We also explored effects of gender, condition, and age, using a mixed-design ANOVA with gender or condition as a between-subject variable and time as a within-subject variable, and a linear regression analysis for age and change in anthropomorphism scores. Third, we used Pearson's correlations to investigate how anthropomorphism and knowledge of L2 words are related (RQ3). We correlated children's scores on the anthropomorphism questionnaire before and after the tutoring sessions, change in scores on the anthropomorphism questionnaire, and scores on the comprehension test on each post-test (i.e., immediate and delayed).

5.3 Results

5.3.1 Anthropomorphism before tutoring sessions

We investigated our first research question: Are there individual differences in the degree that children anthropomorphize the robot? Table 5.2 displays the questions of the questionnaire and the proportions of children that answered the question with "yes". As a group, children tended to attribute more human-like properties to the robot than machine-like properties as is reflected in the overall proportions being higher than .50 at both before and after the tutoring sessions, but the scores varied strongly between the questions. Children highly agreed that the robot "can enjoy something", "can be happy", and "can think". They disagreed more often on various biological properties, such as "Do you think Robin the robot feels it when you tickle Robin the robot?" and "Do you think that Robin the robot can feel pain?".

We explored whether there were effects of gender, age, and condition. The mean anthropomorphism scores, separated for gender, age, and condition, are displayed in Table 5.3. An independent-samples t-test showed no effect of gender, $t(102) = -.30, p = .77, d = .06$, and a linear regression analysis showed no effect of age, $F(1, 102) = 2.24, p = .14$. With respect to condition, we explored whether children perceived the robot differently in the iconic-gesture condition compared to the condition without iconic gestures as measured before the robot interaction, using an independent-samples t-test. There were no differences between the two conditions in the degree to which children anthropomorphized the robot, $t(102) = -.36, p = .72, d = .07$.

Table 5.2: Proportions of Children Answering Yes on the Questionnaire before and after the Tutoring Sessions.

Do you think that Robin the robot...	Before	After
... can see things?	.79 (82)	.81 (84)
... can be sad?	.66 (69)	.41 (43)
... remember something?	.64 (67)	.69 (72)
... can feel it when you tickle Robin the robot?	.45 (47)	.33 (34)
... can think?	.78 (81)	.65 (68)
... has to eat?	.27 (28)	.17 (18)
... understands when you say something?	.66 (69)	.74 (77)
... can feel pain?	.46 (48)	.29 (30)
... can enjoy something?	.92 (96)	.92 (96)
... grows?	.15 (16)	.12 (12)
... can be happy?	.94 (98)	.87 (90)
... can recognize you?	.49 (51)	.89 (92)
Overall scores	.60 (60)	.57 (60)

Note: The total number of children can be found between brackets.

5.3.2 Change in anthropomorphism after tutoring sessions

Then, we investigated our second research question: How does the degree to which children's anthropomorphize the robot change through multiple L2 tutoring sessions with the robot? There was a positive and moderately strong correlation between scores before the tutoring sessions and after the tutoring sessions on the anthropomorphism questionnaire, $r(104) = .505, p < .001$, indicating moderate overall stability of anthropomorphism. However, there was also large variability among the children in whether and how the degree to which they anthropomorphized the robot changed before and after the tutoring sessions. Most children were consistent in the degree to which they anthropomorphized the robot (45 children), that is, their anthropomorphism scores during the two test moments were the same or differed by a maximum of one question. However, a relatively large number of children anthropomorphized the robot less after having interacted with it in the tutoring sessions (35 children). An increase in anthropomorphism also occurred, but was least common (24 children). We compared children's answers on the anthropomorphism questionnaire after the tutoring sessions to their answers before the tutoring sessions. Table 5.2 shows that children changed their opinion drastically on a number of questions. Fewer children believed after the

Table 5.3: Children's mean anthropomorphism scores (SD) before and after the tutoring Sessions, separated for gender, age, and condition.

		Before	After
Gender	Male	.60 (.20)	.53 (.22)
	Female	.61 (.19)	.62 (.17)
Age	Younger	.62 (.19)	.56 (.20)
	Older	.59 (.20)	.59 (.20)
Condition	No iconic gestures	.60 (.19)	.59 (.20)
	Iconic gestures	.61 (.20)	.57 (.20)

tutoring sessions that the robot could feel it when being tickled, that it could feel pain, or that it could be sad. More children believed after the tutoring sessions that the robot could understand what they said, and that the robot could recognize them. However, a paired samples t-test did not show significant differences between children's overall scores before and after the tutoring sessions on the anthropomorphism questionnaire, $t(103) = 1.53, p = .13, d = .15$. We then explored whether there were effects of gender, age, and condition (see Table 5.3 for the means). A mixed-design ANOVA with gender as a between-subject variable and test moment (before and after the tutoring sessions) as a within-subject variable showed an interaction between gender and test moment, $F(1, 102) = 4.35, p = .04, \eta p^2 = .04$. Boys assigned more human-like qualities to the robot before interacting with the robot than after the tutoring sessions, $t(54) = 2.28, p = .03, d = .32$, while there was no difference in girls' anthropomorphism scores between the two test moments, $t(48) = -.54, p = .60, d = .07$. The interaction is displayed in Figure 5.4. Moreover, an interaction effect with age was found. A linear regression analysis was used to predict the difference score in anthropomorphism from age. Age significantly predicted the change in anthropomorphism over time; $F(1, 102) = 5.56, p = .02$, with an R of .05. Children's predicted changed anthropomorphism score is equal to $-0.68 + 0.01 \times$(age in months). Figure 5.5 shows that a younger age was associated with a larger decrease in anthropomorphism before to after the tutoring sessions. Participants' change in anthropomorphism increased 0.01 for each month of age.

To explore whether children perceived the robot differently in the iconic-gesture condition compared to the condition without iconic gestures over time, a mixed-design ANOVA with condition as a between-subject variable and test moment as a within-subject variable revealed that condition did not interact with time, $F(1, 102) = .64, p = .43, \eta p^2 = .01$. Thus,

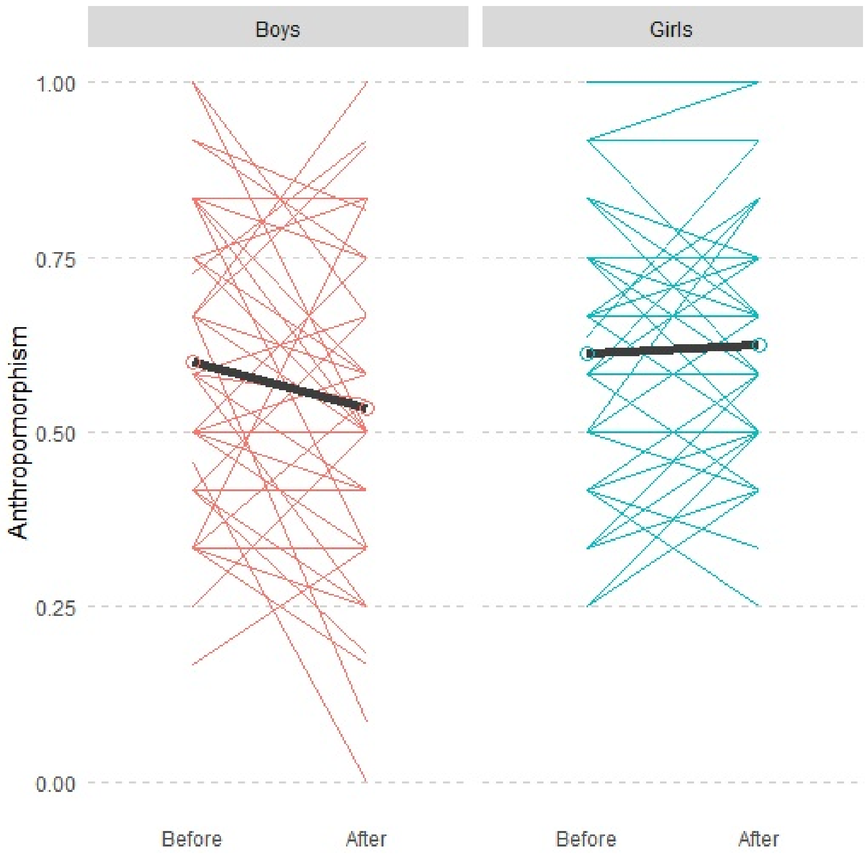


Figure 5.4: Anthropomorphism scores as a function of gender before and after the tutoring sessions.

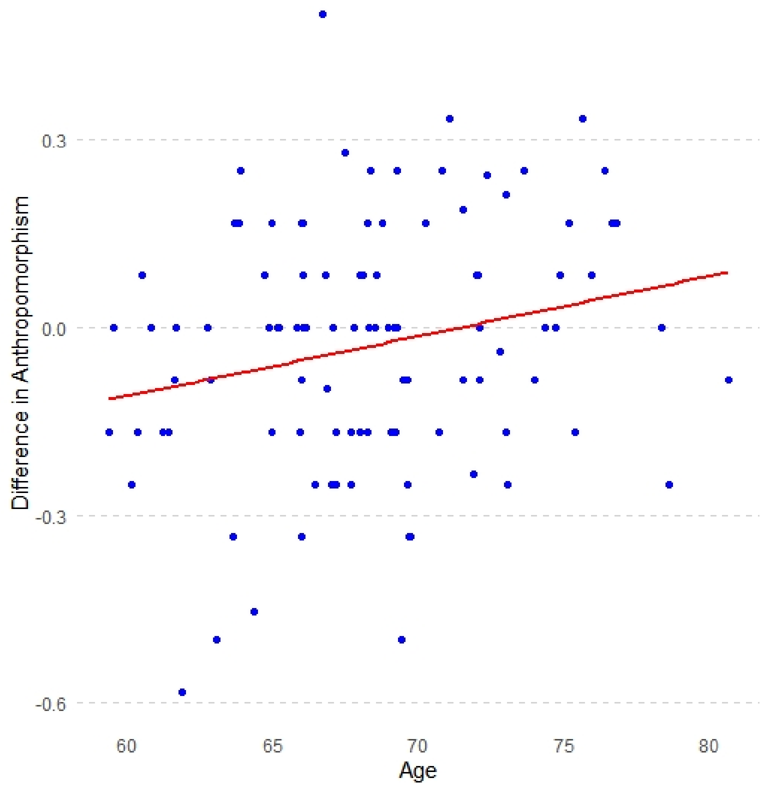


Figure 5.5: Age and the difference in anthropomorphism scores.

the use of iconic gestures apparently was not associated with a different change of children's anthropomorphizing of the robot.

5.3.3 Anthropomorphism and word knowledge

Finally, we investigated our third research question: How are children's anthropomorphistic perceptions of the robot and their knowledge of L2 words related? As already mentioned, we only included the children's comprehension test scores to look at the relation with anthropomorphism. Table 5.4 displays children's scores on the comprehension test during both post-tests.

Pearson's correlations showed that anthropomorphism before the tutoring sessions was weakly related to the comprehension scores on the immediate post-test, $r(104) = -.208, p = .03$ (see Table 5.5). The relation was negative, suggesting that children who anthropomorphized the robot more prior to the session series knew fewer words during the immediate

Table 5.4: Children's mean scores (SD) on the comprehension test.

Condition	Immediate post-test	Delayed post-test
Iconic gestures	29.47 (5.85)	30.43 (6.22)
No iconic gestures	29.39 (6.08)	29.75 (6.44)

Note: The maximum score on the comprehension test was 54.

post-test than children who anthropomorphized the robot less. Anthropomorphism after the tutoring sessions was not related to comprehension scores on either post-test, both ps $> .090$. Children's change in anthropomorphism was weakly but significantly and positively related to the comprehension scores on the delayed post-test, $r(104) = .212, p = .03$. Thus, the larger the change towards anthropomorphism of the robot over time, the higher the performance on the delayed post-test and vice versa.

Table 5.5: Correlations between the anthropomorphism scores and the L2 comprehension scores.

	Comprehension	
Anthropomorphism	Immediate post-test	Delayed post-test
Before tutoring sessions	-.208*	-.137
After tutoring sessions	-.167	.074
Change	.036	.212*

Note: Statistical significance: * p < .05.

5.4 Discussion

In this chapter, we investigated (1) the degree to which five-year-old children anthropomorphized a social robot, (2) whether the degree of their anthropomorphism changed after intensive experience with the robot acting as a peer tutor in an L2 word learning intervention, and (3) whether anthropomorphism and the change therein were related to children's word knowledge.

5.4.1 Anthropomorphism before tutoring sessions

We investigated the way children perceived the robot after a group-wise introduction session and prior to the tutoring sessions. Overall, children slightly more often agreed than disagreed with statements attributing human-like properties to the robot, but there were large differences between children in the degree to which they anthropomorphized the robot, in line with research on individual differences in the tendency to anthropomorphize objects (Epley et al., 2007, 2008). Moreover, children agreed more often with statements that attributed cognitive and, to some extent, also positive emotional states to the robot than biological properties and negative emotional states, in line with previous work that also found that young children are likely to ascribe cognitive mental states to robots (Beran et al., 2011). As this was not the scope of the current chapter, we did not present and analyze children's answers to the open-ended questions, which asked them to motivate why they perceived the robot as more or less human-like. However, we noticed that there were large differences between the children, similar to their overall anthropomorphism scores, in the way they explained why they perceived the robot in the way they did. For example, some children thought that the robot would be sad if children did not want to play with it, while other children thought the robot would be sad if it was in pain. Some children thought that the robot could not be sad because it had no feelings, while other children thought the robot could not be sad because it could not handle water and, thus, could not cry. Contrary to our expectation that gestures would make the robot more human-like, children did not anthropomorphize the robot more when it used iconic gestures than when it only used deictic gestures. This might be due to our design of the experiment, as the robot used the same repetitive gesture each time it used a target word. This repetitive behavior could have reduced the positive effect of the gestures in respect to human-likeness of the robot. As humans do not use the same gesture each time they use a target word, variation in gesture use might increase the human-likeness of the robot again, which would be interesting to explore further. It is possible that in this study the iconic gestures did not convey the concepts as clearly as they were intended and as a consequence of that, the gestures did not impact anthropomorphism. The robot used iconic gestures for each target word and some words were more difficult to act out, such as 'more', for which an iconic gesture is not that iconic. This is supported by the lack of differences in learning outcomes (Vogt et al., 2019), in contrast to a different study (de Wit et al., 2018), where the iconic gestures clearly portrayed the meaning of the word and in which differences in word knowledge were found.

5.4.2 Change in anthropomorphism after tutoring sessions

We investigated the degree to which children anthropomorphized the robot had changed after the L2 tutoring sessions. There were no significant differences in overall anthropomorphism, and, similar to the pre-test, children on average slightly more agreed than disagreed with attributing human-like properties to the robot at the post-test. However, with regard to specific properties some major changes were observed. Fewer children answered "yes" to questions attributing biological properties and negative emotions to the robot at the post-test as compared to the pre-test. This concerned, for example, questions asking whether the robot "could feel it when being tickled" or "could feel pain". This is in line with the study of Sciutti et al. (2014), who found that the robot's sensory and motor properties became more salient to children after they had interacted with a robot. At the post-test, more children answered "yes" to questions addressing cognitive abilities, such as whether the robot can remember something, understand them when they say something, and is able to recognize them. These changes together indicate an interesting shift in the way in which the robot is seen by children after intensive experience, namely as a basically mechanical being but with positive mental states, whereas initially children showed more confusion regarding the biological aspects and were less strongly convinced of the cognitive capabilities of the robot. We believe that this shift is due to the way in which the sessions were designed. At the start of each session, the robot greeted the children personally while mentioning their names, referred to the previous sessions and tracked the children's faces to suggest that the robot looked at them. The open-ended answers confirmed that possibility as their explanations changed from "Robin the robot has not met me yet" to "Robin the robot said my name every time we played". It is likely that children were less inclined to believe that the robot could recognize them at the pre-test, simply because they had not yet played intensively with the robot in a one-on-one setting yet at that time. The same shift was found in the explanations of children for the question whether the robot can remember something: children started with many different explanations before the interaction like "No, Robin the robot has small ears so cannot remember much", "Yes, Robin the robot looks like a human so can also remember things" and changed their explanation after the interaction to "Yes because Robin the robot remembered where we played before". Regarding negative emotional states, fewer children believed at the post-test that the robot "could be sad", which can also be explained by the design of the sessions. Even though the robot expressed happiness (by changing the colors of its eyes) and also when it was not specifically happy (by not changing the colors of its eyes), it never expressed negative emotions, such as sadness or anger. Again, this was supported by the children's open answers where they mentioned the robot's colored eyes during the post-test questionnaire. Most children anthropomorphized the robot either to the same degree or to a lesser degree during the post-test as compared to the pre-test. Fewer children increased their anthropomorphism of the robot. Explorative analyses showed that age and gender had an influence on the change in anthropomorphism: Boys and younger children had a larger decrease in anthropomorphism than girls and older children. It is possible that decreases in anthropomorphism were due to children initially having high expectations of the robot's interactive (human-like) qualities, which the robot could not meet (Dautenhahn & Werry, 2004). This effect could have affected the younger children more as older children seem to anthropomorphize robots less in general (van Straten, Peter, Kühne, & Barco, 2020). Moreover, gender influenced the change in anthropomorphism. It is possible that girls were more forgiving of the robot's flaws than boys were and that girls, therefore, did not change their perception as much as boys (Tung, 2011).

The robot was semi autonomous during the tutoring sessions, but did not engage in personalized conversations with the children. The robot kept to the preprogrammed script and did not answer children's questions. For children with high expectations regarding the human-likeness of the robot, this could have led them to decrease their attribution of human-like properties to the robot. Conversely, children who had a less human-like perception of the robot prior to the tutoring sessions may have had low expectations of the robot's interactive (human-like) qualities. Since the robot displayed at least some human-like behaviors, such as mentioning the child by name (suggesting that it recognized the child) or indicating that it liked the sessions, this could have increased children's beliefs about the robot as human-like over repeated interactions. Thus, the observed changes in anthropomorphism may not only have been dependent on the robot's behaviors (in line with Tung (2016)), but also on whether this behavior matched children's prior expectations. A final possibility is that the observed change in anthropomorphism merely reflects the phenomenon of regression to the mean, with initially higher scores decreasing and initially lower scores increasing at post-test due to random measurement error. While we cannot fully rule out this explanation, it should be noted that more children decreased rather than increased in anthropomorphism, and the analysis at the item level revealed a complex but interpretable pattern of changes that pointed to a shift in how children perceived the robot within a similar overall anthropomorphism score at the pre-test and post-test.

5.4.3 Anthropomorphism and word knowledge

Finally, we investigated whether anthropomorphism and word knowledge were related. We found two weak but significant correlations. Children's anthropomorphism of the robot at pre-test was negatively related to their comprehension scores at the immediate post-test, though not at the delayed post-test. In contrast, a change in perception towards more anthropomorphism was positively related to word knowledge at the delayed post-test, though not at the immediate post-test. Possibly, both correlations point again to the role of children's expectations about the robot as a human-like being. If children had low expectations of the robot and the robot exceeded these expectations, they may have become more engaged, which is beneficial for learning. In contrast, children with high expectations which the robot could not meet, may have become disappointed while working with the robot over several tutoring sessions. There are two important caveats concerning this link between anthropomorphism and word knowledge. First, the correlations, though statistically significant, were rather weak. Moreover, we did not include child characteristics such as age and cognitive ability that could possibly underlie the observed correlations. It is possible that the correlations are spurious and caused by a shared third factor. Second, the present design did not allow for testing the causal direction of the observed correlations. Thus, it is not clear whether children learn more from the robot because they come to perceive it more as human-like, or that they come to perceive the robot as more human-like because they have successful language-learning interactions with it.

5.4.4 Limitations, strengths, and future research

The current study has several limitations. First, we did not use a standardized questionnaire for anthropomorphism because of our young target group. Standardized tests, such as the Godspeed questionnaire (Bartneck et al., 2009), often use Likert scales or semantic differentials, which are too difficult for young children. In contrast, other measures that are specifically designed for young children and are therefore more appropriate to use (Kory-Westlund & Breazeal, 2019a), do not capture the type of human-like properties children attribute to robots. We based our questionnaire on previous work (Jipson & Gelman, 2007) and the questionnaire was found to be reliable, showing also moderate stability between pre-test and post-test. The proposed questionnaire can therefore be seen as a first step towards a validated questionnaire to measure children's anthropomorphism of robots. Furthermore, we do not know how the introduction of the robot before the pre-test affected the degree to which children anthropomorphized the robot. To ensure that children could establish a common ground with the robot and to decrease any anxiety, the introduction contained several statements about the properties of the robot that related to, amongst others the robot being a peer, speaking as a human and looking as a human. It is possible that these statements may have biased children's perception towards anthropomorphism at the pre-test. However, administering the anthropomorphism questionnaire prior to the introduction would have had other disadvantages. For instance, it would not have been clear whether children's perceptions were based on actual interactions with similar robots, with different robots, or were based on cartoons, movies or television programs, or just on imagination. The large variation in scores indicates that children still formed their own opinions about the robot, but we do not know whether these opinions were biased towards anthropomorphizing. Note that despite this possible bias, the changes in anthropomorphism we observed, in particular at the item level, can be considered genuine and likely to relate to the intensive experience children had with the robot during the sessions. Moreover, we could only conduct correlational analyses to examine how anthropomorphism and word knowledge were related. Moreover, we could not rule out that other child-related factors underlie the relations that were observed between children's anthropomorphism and word knowledge. Future research with field experiments is needed to test whether framing the robot as a machine or as similar to a human affects children's learning differently. A high level of anthropomorphism in itself may not be required for successful tutoring sessions, as no positive main effects of anthropomorphism were found in our study. On the other hand, managing children's expectations of robots especially at first, may be important, as lower initial levels, indicating more reserved expectations of the robot, relate to more word knowledge than when expectations are (too) high. Furthermore, it is difficult to translate these results to other fields in which technology is used to support learning, such as VR, AR, XR or serious games. These types of technology often use virtual avatars, which users may anthropomorphize and may thus be subject to similar relations between anthropomorphism and learning outcomes as in our study. It is possible that, since no differences could be found between the two different robot conditions, interacting with a robot over a longer period of time is more important for children's anthropomorphism than specific behaviors the robot displays, such as gestures. Such behaviors of the robot can still be important for anthropomorphism, but mainly in short interactions (Tung, 2016) and after multiple exposures, the interaction itself becomes more important (e.g., the conversations or type of activity that the child and robot engage in). This would give an indication that our results can also be translated to other fields. However, as we only measured children's perception with a robot, we will need to investigate more thoroughly to determine whether this is the case. Finally, there are studies suggesting that presenting robots as human-like to children is undesirable (Broadbent, 2017), as a robot expressing simulated feelings as real feelings is deceptive. Moreover, children may form relationships with robots that may come at the cost of relationships with people. It is important for developers to make sure that children realize that robots are different from human beings. Repeated exposure may more easily reveal a robot's flaws and thus lead to decreases in anthropomorphism, but a subset of the children in our study were found to increase in anthropomorphism, despite our robot's flaws. This is in line with a study finding higher anthropomorphism after repeated exposure (Michaelis & Mutlu, 2018). Thus, even after engaging with a "flawed" robot, children may continue to anthropomorphize a robot. Therefore, researchers may want to consider whether presenting the robot as a social entity and suggesting it has cognitive, emotional, or social abilities is required for their study. After all, even though transparency about the robot's lack of psychological abilities leads to lower anthropomorphism, children may feel as close to the robot as when children's expectations about the robot's psychological abilities are managed (van Straten, Peter, Kühne, & Barco, 2020). Our study also has several strengths. It is one of the first studies to investigate anthropomorphism and changes therein after children had multiple interactions with a robot, and to relate it to children's word knowledge. Furthermore, we included a large sample of young children. Finally, the different robot properties presented in the questionnaire allowed for a more thorough and differentiated understanding of the ways in which children perceive robots.

5.5 Conclusion

The study presented in this chapter explored the degree to which children anthropomorphize a social robot, whether this had changed after seven tutoring sessions, and whether anthropomorphism correlated with children's word knowledge after these sessions. We found that children generally anthropomorphized the robot, although there were large differences between children in the degree to which they did. Our results showed that children's overall tendency to anthropomorphize had not significantly changed after the tutoring sessions, but the analysis at the item level revealed a complex pattern of changes indicating a shift within this overall tendency towards seeing the robot as more mechanical while at the same time attributing more cognitive capabilities to the robot. As an exploration, we found a weak but significant correlation between children's increased anthropomorphism and their word knowledge. Children who came to perceive the robot more as a human knew more words after the tutoring sessions. Although the causal direction of this relation is not yet clear, the results underscore the importance of taking children's anthropomorphism into consideration when designing robot-assisted tutoring sessions.

6

Discussion

Engagement is important within second-language learning. When children remain engaged, they are arguably also more motivated and they will actively focus more on the task which can potentially result in better *short-term* learning gains (Morgan et al., 1990) and *long-term* learning gains (Dörnyei, 1998). However, the evidence that engagement contributes to learning in child-robot interaction is still sparse. Therefore, the main focus in this dissertation was engagement in child-robot interactions.

This dissertation reported multiple studies in which children interacted with a social robot. In these studies, the robot acted as a tutor teaching children different second-language words. The study in Chapter 2 investigated the effect of peer-like and adult-like feedback on children's task engagement, robot engagement and word knowledge and zoomed in on the role of gaze in engagement. Chapter 3 described a study that based the feedback on strategies recommended by student teachers and investigated the effect of this feedback over time on children's task engagement, robot engagement and word knowledge. Chapters 4 and 5 both reported on one study, a study that examined the robot's presence on task engagement, the robot's use of iconic gestures on children's task engagement and robot engagement (Chapter 4) and children's perception of the robot (Chapter 5). The experiments showed that the behavior of the robot (feedback and gestures) had an effect on children's task engagement and robot engagement and children's engagement scores and their perception of the robot vary over time. Moreover, while children anthropomorphized the robot, its non-verbal behavior did not have an influence on their perception. Finally, our results showed that children can learn from the robot, but that the robot is not necessarily better than a tablet in teaching L2 vocabulary.

6.1 Insights in measuring engagement

There are several aspects of a learner's behavior that are important for determining their engagement, such as their eye gaze, posture and emotional expressions. Of these aspects especially eye gaze can play a large role because this shows the direction of the learner's attention. Accordingly, we first identified the role of eye gaze, as specified in

> **Research question 1**: Can children's eye gaze be used to monitor their task engagement and robot engagement?

In Chapter 2, we showed that both task engagement and robot engagement could largely be predicted by the duration that children looked in a certain direction. Regression analyses revealed that, assuming a default high *task* engagement level (i.e. a relatively high intercept), gaze toward targets involved in the interaction reduced the engagement level. This happened more so for gaze toward the experimenter and elsewhere than toward the robot and the blocks (i.e. the negative slopes for gazing toward the robot and blocks were half the size than for looking at the experimenter and elsewhere). This suggests that disengagement is easier to detect (by gazing away) than engagement (by gaze in direction of the task). *Robot* engagement was also negatively affected by gaze directions elsewhere and to the experimenter but in contrast to task engagement it was positively influenced by children's eye gaze toward the robot, suggesting that gaze toward the social partner is more important for robot engagement.

These results were consistent with previous studies although these researchers did not specify the type of engagement they were measuring. Ishii et al. (2013) used eye gaze to determine disengagement with adults, found a correlation between eye gaze and disengagement, and used this finding to regain participants' engagement by providing re-engaging behavior by the robot. Moreover, when interviewing preschool teachers about engagement indicators, eye contact was an indicator for engagement and gaze away was an indicator for disengagement (Schodde et al., 2017). It is important to note that we measured eye gaze during one specific episode of two minutes. This shows that eye gaze was related to engagement over a longer period during the interaction. However, this duration might also have a limiting effect, because we did not measure eye gaze during short periods meaning that if engagement was low for a shorter duration, we only measured the overall effect.

These differences between the role of eye gaze on task engagement and on robot engagement highlight that it is important to differentiate between children's engagement with the task, and engagement with the robot. This is because the distinction can clarify whether children are more involved with the task or the robot's behavior when executing a robot-tutoring learning task. Likewise, distinguishing the two engagement types can show the effect of the robot's presence as a social partner. After all, children may consider the robot's behavior as interesting, and as a result be engaged, but may not like the task or vice versa. Therefore, when researchers separate these two concepts, it provides more insight regarding children's engagement and the factors influencing that. Moreover, if we separate the two, we can observe the effect of both separately on children's learning outcomes.

However, both engagement types could not completely be predicted by children's eye-gaze directions which demonstrates that despite eye gaze having a large role in engagement, it does not predict all aspects of engagement. There is still a possibility that a child is looking at the robot during the whole interaction while not being robot-engaged and vice versa. As preschool teachers indicated (Schodde et al., 2017), head movements such as nodding and shaking, resting their head on their hand, movements, answering the questions are potential additional factors that could be taken into account. In fact, Perugia et al. (2020) tried to model engagement based on attention measures, valence and arousal and concluded that it is impossible to measure them separately. Therefore, the studies in later chapters of this thesis did not rely on analyses of eye gaze, but rather relied on a coding scheme that included other aspects of engagement, such as children's expressions, on top of eye gaze. This coding scheme also distinguishes between engagement with the robot, and engagement with the task in order to get a complete view of engagement.

Finally, not only the *learner's* gaze behavior is important but also the *robot's* gaze behavior. The large role of eye gaze for determining engagement can also be interpreted in a way that the *robot* should use correct gaze behavior in order to show the learner that the robot is engaged in the interaction. Gaze behavior by the robot has indeed been assumed to be important for the interaction and has been used in earlier work to provide cues during the interaction and as a result can influence engagement (Mwangi et al., 2018). In this dissertation we did not look at gaze behavior of the robot specifically, however we did investigate the effect of other non-verbal robot behaviors on engagement, such as iconic gestures.

6.2 Factors that influence engagement

With insights into how we can measure children's task engagement and robot engagement, we investigated what kind of robot behavior can influence these two engagement types. We concentrated on two factors that can benefit children's learning and engagement, answering

> **Research question 2:** Do robotic feedback and iconic gestures influence children's task engagement and robot engagement?

We will first discuss the influence of feedback on engagement and then of gestures.

6.2.1 Feedback

Our studies in Chapters 2 and 3 revealed that type of feedback provided by the robot has an effect on task engagement and robot engagement (although not on learning gains). These chapters showed that feedback makes tasks encouraging and engaging. Chapter 2 showed that three-year-old children's task engagement and robot engagement tended to increase during the lesson when receiving feedback, whereas without feedback children's engagement tended to decrease over time. Consequently, we suspected that feedback would have more impact when provided over a longer time period and in Chapter 3 we therefore explored the effect of feedback over the duration of three different sessions. Moreover, we wanted to make sure that we designed the feedback to be as effective as possible and, thus, we interviewed student teachers how they would provide feedback using the robot and implemented their answers into a teacher-preferred feedback strategy and teacher-dispreferred feedback strategy.

In Chapter 3, we implemented these strategies onto the robot and found that children aged 5 and 6 scored higher on task engagement and robot engagement when the robot provided teacher-preferred feedback than when the robot did not provide any feedback or feedback dispreferred by teachers. Furthermore, we found that the order in which the different types of feedback were provided had an influence on children's task engagement: children who first received teacher-preferred feedback and then a different type, scored lower on task engagement in the following sessions.

Since the two studies used different types of feedback (with children of different ages) it is not possible to directly compare them, but both studies showed that feedback tends to have a positive effect on children's engagement. This finding is consistent with data obtained in human-human studies in which positive feedback increases children's motivation (Blumenfeld et al., 2006). Moreover, negative feedback can result in a higher task confidence because children can correct themselves after receiving this negative feedback which can result in a higher engagement. Our studies have been unable to demonstrate whether positive or negative feedback had different effects. Based on our data, we tentatively conclude that the role of different *kinds* of feedback is small, but that feedback itself is important for children's engagement.

6.2.2 Gestures

In Chapter 4, we investigated the effect of the robot's iconic gestures on children's task engagement and robot engagement. Robot gestures indeed had an influence on children's task engagement and robot engagement, but in an opposite pattern from what we expected. *Task* engagement was generally higher for children when interacting with a robot using no iconic gestures than with a tablet or with a robot using iconic gestures. The latter only holds when comparing the first three sessions with the following three sessions. *Robot* engagement was higher for children interacting with a robot using iconic gestures than without iconic gestures.

These results can possibly be explained by looking closer at the gestures of the robot. Both robot conditions used deictic gestures to redirect the child's attention to the tablet when the child had to execute a task on the tablet. These deictic gestures could have resulted in a higher task engagement for both condition, however we only saw a higher task engagement for the no iconic gesture condition. This inconsistency, therefore, is likely caused by the iconic gestures used in the iconic gesture condition. These iconic gestures might have attracted the child's attention to the robot when there was a task on the tablet, resulting in a *lower* task engagement, as suggested by Kennedy et al. (2016). This is, moreover, extra important for certain groups of children. In a study on learning math tables with a robot (Konijn & Hoorn, 2020), children who had a lower school performance were more easily distracted than children with a higher school performance and showed to be more distracted by social behavior of the robot.

A similar reasoning can be provided for the results of robot engagement. A robot in the iconic gesture condition that moves its arms and hands will draw more attention to itself, resulting in a *higher* robot engagement. Moreover, a robot using no iconic gestures will draw less attention to itself than a robot using iconic gestures, resulting in a *lower* robot engagement. These findings are consistent with our RQ1 results, where we found that robot engagement had a positive relation with children's gaze toward the robot. When the robot draws more attention to itself by using iconic gestures, this seems to influence robot engagement positively.

These findings imply that the robot's gestures must be well designed in an interaction. For example, when robot engagement is important for the task, such as when the robot is actively having a conversation with the child, (iconic) gestures can be used to draw attention to the robot. However, when there is a task on a tablet that needs attention, it may be more helpful not to use (distracting) gestures and only provide deictic gestures to redirect the child's attention to the tablet.

Finally, when combining these two robot behaviors: feedback and gestures, it is possible that the non-verbal behavior during the feedback played a role in engagement. The robot used non-verbal behavior when providing the engagement-increasing feedback. For instance, the robot nodded when providing positive feedback and used colored eyes to indicate happiness. Previous studies showed that motivational gestures, such as a high five, thumbs up or a fist bump played a role in engagement (van Minkelen et al., 2020; Morris & Zentall, 2014) and that children paid most attention to feedback accompanied by an arm gesture (Serholt & Barendregt, 2016). Future research should explore this role of motivational gestures.

6.3 Novelty effect

Engagement is not only dependent on the robot's behavior, it also is dependent on time. To investigate the novelty effect of the robot and the effect of multiple tutoring sessions, we specified the following

> **Research question 3:** How do children's task engagement and robot engagement develop over time?

For this purpose, we measured children's interactions over multiple sessions in Chapters 3 and 4. Both studies showed that children's engagement was relatively high during the first interaction and that it decreased over time, in accordance with the novelty effect, but that there were large individual differences between children. The experiment in Chapter 3 included three sessions, and showed that overall, children's task engagement and robot engagement tended to decrease during these three lessons. This experiment was a within design, and children received three types of feedback: teacher-preferred, teacher-dispreferred and no feedback in a random order. However, it seemed that children's task engagement was influenced by the robot's feedback behavior in the previous lesson. In the case that the children first received teacher-preferred feedback, and then another form of feedback (no feedback or teacher dispreferred), their task engagement decreased significantly. Receiving no feedback or teacher-dispreferred feedback after a session with the robot using teacher-preferred feedback might have demotivated the children and as a consequence decreased their engagement. It is therefore important to take previous interactions into account when designing long-term child-robot interactions (Leite, Martinho, & Paiva, 2013).

When comparing task engagement with robot engagement over time, we found no large differences. Both engagement types are influenced by this novelty effect and the individual differences. This means that researchers have to take into account that both engagement types can decrease over time.

In Chapter 4, we compared the two engagement types over more sessions. This study included a group introduction, six tutoring lessons and one recap session, which meant that children saw the robot twice a week during a full month. When examining children's engagement patterns, we noticed that after the third lesson, children's engagement seemed to decrease less (although it still varied for each child and lesson). Interestingly, this seems to be consistent with the suggestion by Salter et al. (2004) that when wanting to examine beyond the novelty effect, researchers should carry out interactions including more than two sessions.

A note of caution is due here since it is difficult to generalize this guideline to other studies because these studies might have a different design. The precise moment when the novelty effect starts to wear off is dependent on the number of interactions in total, the length of these interactions and the behavior of the robot (Leite, Martinho, & Paiva, 2013).

We found that children's task engagement and robot engagement increased again during the final session, the recap session, possibly because the interaction during the recap session was different from the other sessions. Introducing novel behaviors and tasks has already been recommended in the past by Leite, Martinho, & Paiva (2013) and been used by Tanaka et al. (2007) to increase children's engagement. Furthermore, it is also possible that because this recap session was more interactive than the other sessions, children became more engaged. For each of the 34 target words, children could drag a virtual sticker into a virtual picture book and they had to repeat the robot after the target word was placed in the picture book, which made this interaction more interactive. This interactivity might have made the children feel more in control, increasing their sense of autonomy which has been previously linked with more motivation (Deci & Ryan, 1985) and with that might increase engagement. Moreover, it is possible that children recognized the words which made the session easier. Finally, it is possible that children knew it would be the last time they would play with the robot and tablet and this might make them attentive during the interaction, making them enjoy it more and hence increasing their engagement.

Both studies in Chapters 3 and 4 showed that there were large individual differences in children's engagement patterns between each session. These patterns varied from a decrease of children's engagement and an increase of children's engagement to more complex patterns such as children being engaged in the first session, not engaged during the second lesson and engaged again in the last session. It is very important to investigate these individual differences more closely and measure engagement continuously during the session in order to personalize the lesson when a child is less engaged by e.g. introducing more breaks during the session or changes in robot behavior (Kim et al., 2020). Konijn & Hoorn (2020) found that children who performed worst at school were easily distracted by the robot's social behavior and suggested that these groups of children might benefit more from neutral behavior by the robot. These findings, showing that the robot's behavior sometimes is an advantage to children but sometimes a disadvantage to other groups of children are a strong argument for personalized lessons.

6.4 Relation between children's engagement and word learning

Human studies indicate that engagement is critical in language acquisition and children learn better when they are engaged (Deci & Ryan, 1985). Therefore, we addressed

> **Research question 4:** What is the relation between children's task engagement and robot engagement, and their second-language learning gain?

In all studies presented in the Chapters 2,3 and 4 we compared children's engagement with children's L2 word knowledge before and after the lessons. The two studies in Chapters 2 and 3 did not show a relation between the two engagement types and word knowledge, but the long-term study in Chapter 4 showed a positive relation between both task engagement and robot engagement and L2 word knowledge. This suggests that both engagement types play a role in word learning over more sessions. It seems possible that these results are due to having more sessions in Chapter 4 compared to the one session and three sessions experiments in Chapters 2 and 3 respectively. A higher engagement might be especially important over multiple sessions, helping children to stay motivated beyond the novelty effect. An alternative explanation could be that children were presented with more L2 words in the long-term interaction and therefore could also learn more words, which might have created larger differences between the highly engaged and not engaged children, resulting in a significant correlation. This suggests that it is even more important that children's engagement remains high over time and interactions are designed to keep this engagement high.

The positive relation between robot engagement and word learning gain is somewhat unexpected, because the results from Kennedy et al. (2015) and Konijn et al. (2021) suggested that the robot's behavior might distract children from the task and we therefore expected that a high robot engagement would decrease children's L2 learning gain compared to a high task engagement. However, in our long-term study we found a positive correlation between children's robot engagement and their word knowledge. It is possible that in our experiment, the robot's behavior was related to the task (because it used iconic gestures for the L2 words) and that children therefore still learned the connection between the L1 concept and L2 concept. This suggests that not all robot attention is bad, if the robot's behavior is related to the task, it might be even better for the children's learning gain.

6.5 Children's perception of the robot

Lastly, we determined whether children's perception of the robot changed over time. When children are interacting with a social robot, they will undeniably develop a social bond with this robot (van Straten, Peter, Kühne, & Barco, 2020). How this perception will change over time, is something we addressed with the following research question:

> **Research question 5:** How do children's perceptions of the robot develop over time when interacting with a robot tutor and is it related to their L2 learning gain?

Chapter 5 shows that on average children's anthropomorphism remained approximately the same before and after seven tutoring sessions with the robot. When looking more closely to the children's answers, we found that children changed their perception mainly regarding the cognitive aspects of the robot. After the seven tutoring sessions, they perceived the robot more as a mechanical being with positive mental states.

This shift can be explained by our design of the sessions. At the beginning of our sessions, the robot greeted the child with their name and referred back to the previous sessions to create common ground between them which could have led the children to believe that the robot had cognitive capabilities. Furthermore, the children attributed mainly positive emotions to the robot and no negative emotions. This can also be explained by our design because the robot used verbal and non-verbal behavior to express positive emotions, such as happiness, but did not express any negative emotions. The question remains what would have happened to children's perception when the robot would also have expressed negative emotions. It is possible that the effect would have been even stronger because a recent study (Nijssen et al., 2021) showed that children perceived a robot as more human-like when it expresses that it has positive and negative feelings, although in a completely different setting.

The Chapter 5 study also showed many individual differences. For example, when comparing boys and girls, we found that boys anthropomorphized the robot less after the lessons than girls. Moreover, we found a relation between children's age and anthropomorphism, where younger children had a larger decrease in anthropomorphism than older children.

In addition, there was a correlation between children's anthropomorphism score and their word knowledge before the lessons and immediately after the lessons. Children's word knowledge scores at the delayed post-test were related to children's change in anthropomorphism. In other words, children who anthropomorphized the robot more after the lesson remembered more L2 words over time, while children who anthropomorphized the robot less after the lesson than before remembered fewer L2 words. Moreover, children's anthropomorphism before the sessions was related to their immediate word knowledge, but not to their retention. If children had low expectations of the robot and the robot exceeded these expectations, they may have become more engaged, which might have been beneficial for their learning. In contrast, children with high expectations which the robot could not meet, may have become disappointed while working with the robot over several tutoring sessions. Overall, this shows there is a relation between children's perception of the robot and their word learning, however it is not yet clear whether there is a causal relationship between them.

Our findings may also shed some light on the relation between children's perceptions and robot engagement. As said in the previous paragraph, in our experiment, we observed that when children had high yet unmet expectations regarding the robot's behavior, children became more confused by and frustrated with the robot, resulting in less robot engagement (D'Mello & Graesser, 2012). For instance, when a child expected the robot to behave in a human-like manner, such as talking back when they told a story, and the robot did not reply, the child might have thought that the robot did not listen to the child and as a result became less excited to tell more stories or interact in a different manner. This is also in line with previous studies in a meta-analysis, (Blut et al., 2021), whose authors suggest a conceptual model in which perception of the robot can be a mediator for, among others, the robot's likability and positive and negative affect for the robot. However, since we did not specifically investigate the relation between anthropomorphism and engagement, this should be explored more.

6.6 Strengths and limitations

The previous sections discussed the studies we carried out in context of the L2TOR project and answered our research questions. Looking back on our studies, there are several strengths and limitations in these studies. This section will expand on these strengths and limitations and identify directions for future research.

6

6.6.1 Strengths

Our studies have various strengths. First, the study described in Chapters 4 and 5 is among the first long-term studies in the HRI field that included a relatively large sample and that was preregistered. Second, in all studies described in this dissertation, we applied the same engagement coding scheme which makes the results more comparable. This engagement coding scheme was adapted from an existing coding scheme to be applicable for robot interactions.

Third, we differentiated task engagement from robot engagement to study the robot as social partner and the role of the task in child-robot interactions. Finally, we familiarized the children before all experiments so they were used to the robot. Young children should first familiarize themselves with the robot to reduce anxiety (Vogt et al., 2017). This introduction explained what the robot is able to do, how it can move and how children can interact with the robot to set their expectations before the one-on-one interactions with the robot.

6.6.2 Limitations

Our study also has several limitations. First, our interactions were not adaptive. One of the advantages of using a robot is that it can adapt itself and its teaching methods or stories for each child which can help long-term relations (Jacq et al., 2016; Ligthart et al., 2019). In our experiments, the robot's behavior followed a script, with some exceptions such as when the robot had to choose between positive or negative feedback. We chose this design because in this way we could control the manipulations, making it possible to directly compare conditions between children. However, we might have found other results if we would have used adaptive scenarios. Earlier adaptive studies did yield mixed results. One study from the L2TOR project found that adaptivity did not support children's learning gain over one session, but it did have an influence on their engagement: children in the adaptive condition had a smaller decrease in engagement during the session than children who were in a non-adaptive condition (de Wit et al., 2018). In addition, in another study in our project (Schodde et al., 2019) the robot informed the learner about its adaptive decisions, providing the learner with insights on how they could improve. These explanations supported the participants' learning gain, especially for slow learners. A different study showed that slow learners also preferred a robot using extra explanations over one without, while fast learners indicated that a robot using extra explanations was going too slow (Hindriks & Liebens, 2019). In a similar way, Ahmad et al. (2019) used a robot that adapted the feedback in the lessons based on the children's emotional state and found that this robot was successful in sustaining children's engagement. Thus, adaptivity should be explored more.

Second, our studies did not include automatic speech recognition. At the time we started the L2TOR project, speech recognition was not reliable enough (Kennedy et al., 2017). In addition, speech recognition of children's non-native pronunciation is more difficult because there is less training data for state of the art recognition software. However, this speech recognition is important when teaching a second language, especially when focusing on production skills. In our experiments we could only focus on comprehension and translation skills.

Third, in most of our studies, the interaction between child and robot were reliant on the tablet game in front of the robot. The main lesson manipulations were executed on the tablet and the robot supported the children during these tasks verbally and non-verbally. This role of the tablet might have directed the attention of the child away from the robot and the robot's behavior such as feedback and gestures. Therefore, children might have focused less on our manipulations and therefore the effect of these manipulations was smaller than anticipated. A recent study showed that children learned more and were more engaged when training with a social robot without a tablet compared with only a tablet (Konijn et al., 2021) and this indicates that it might be worthwhile to reduce the role of the tablet in follow-up studies.

Fourth, it is difficult to generalize our results to other age groups. Our study only included young children of 3 to 5 years old. It is possible that older children might respond differently to the robot's behavior. In our L2TOR studies, we found that older children often perform better and benefit more from the robot's behavior (e.g., Chapter 4 of this dissertation and de Wit et al., 2020).

Finally, non-native people often rely more on gestures than native people in a language, however they are also bothered more by distracting sounds in the background when the language is non-native (Drijvers et al., 2019). It is not clear yet whether the robot's noise of its motors is actually doing more harm than good during the gestures and whether the effect of gestures is larger than the negative effect of the noise of the motors. In a study by Zhang & de Haas (2020) participants learned the different tones in Chinese, something very reliant on sound. The participants performed better when interacting with the robot without gestures, possibly because the noise of the robot's accompanying gestures that made it harder to hear the specific tones. To reduce this noise effect in our experiment, we designed our sessions taking the motors' noise into account, and making sure that the robot's L2 words were pronounced after the robot used an iconic gesture. However, this resulted into a slowed interaction because children's had to wait until the gesture was finished before hearing the L2 word. While newer robots make less noise when moving their arms, it would be interesting for future studies to take this noise more explicitly into account when designing second-language learning interactions.

6.7 Concluding remarks

This dissertation expands our knowledge of using a robot tutor to teach young children a second language and showed that children are engaged with a robot and task and that over long-term, children's engagement has a positive relation with children's learning gains. The first sentence in this dissertation was about a future image of using a robot in the classroom as support for teachers. This vision has only become more relevant since I started my PhD research because the recent COVID-19 pandemic has once more made it painfully evident how much education depends on digital innovations.

When schools were closed and caregivers had to home-school their children, there were widespread concerns about the impact of the lock-down measures on the quality of education. Digital tools, including social robots and AI, which can adapt to individual learners, are available all the time, have infinite patience, and can be distantly monitored by professional human teachers, would have been extremely valuable in this situation, supporting teachers and schools.

Not only for at-home education during our pandemic, but even after the pandemic this trend may continue. However, before being able to introduce the robot in education and making the future classroom a reality, it is important to take further steps. On the one hand, it is good to already deploy robots for tasks they are currently suited for, such as single lessons tailored to maximize the contribution of robots. On the other hand, for the future, more studies need to be carried out over longer periods and without an experimenter having to control the robot. With these extra experiments in mind, we are confident that robots will increasingly be able to support the teachers of the future.

References

Ahmad, M. I., Mubin, O., & Orlando, J. (2017). Adaptive social robot for sustaining social engagement during long-term children–robot interaction. *International Journal of Human–Computer Interaction*, *33*(12), 943-962. doi: 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. Retrieved from https://doi.org/10.1177/1059712319844182 doi: 10.1177/1059712319844182

Alemi, M., Meghdari, A., Basiri, N. M., & Taheri, A. (2015). The effect of applying humanoid robots as teacher assistants to help iranian autistic pupils learn english as a foreign language. In *International conference on social robotics* (p. 1-10). Springer, Cham.

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

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

Arnott, L., Grogan, D., & Duncan, P. (2016). Lessons from using ipads to understand young children's creativity. *Contemporary Issues in Early Childhood*, *17*(2), 157-173. Retrieved from https://doi.org/10.1177/1463949116633347 doi: 10.1177/1463949116633347

Assad-Uz-Zaman, M., Islam, M. R., Miah, S., & Rahman, M. H. (2019). Nao robot for cooperative rehabilitation training. *Journal of Rehabilitation and Assistive Technologies Engineering*, *6*, 2055668319862151. Retrieved from https://doi.org/10.1177/2055668319862151 (PMID: 31413864) doi: 10.1177/2055668319862151

Axelsson, E. L., Williams, S. E., & Horst, J. S. (2016). The effect of sleep on children's word retention and generalization. *Frontiers in Psychology*, *7*(AUG), 1192. doi: 10.3389/fpsyg.2016.01192

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617-645.

&

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71-81. doi: 10.1007/s12369-008-0001-3

Baxter, P., de Jong, C., Aarts, R., de Haas, M., & Vogt, P. (2017). The effect of age on engagement in preschoolers' child-robot interactions. In *Proceedings of the companion of the 2017 acm/ieee international conference on human-robot interaction* (p. 81–82). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3029798.3038391

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

Belpaeme, T., Vogt, P., van den Berghe, R., Bergmann, K., Göksun, T., de Haas, M., ... others (2018). Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics*, *10*(3), 325-341.

Beran, T. N., Ramirez-Serrano, A., Kuzyk, R., Fior, M., & Nugent, S. (2011). Understanding how children understand robots: Perceived animism in child–robot interaction. *International Journal of Human-Computer Studies*, *69*(7-8), 539-550.

Bernstein, d., & Crowley, K. (2008). Searching for signs of intelligent life: An investigation of young children's beliefs about robot intelligence. *The Journal of the Learning Sciences*, *17*(2), 225-247.

Betts, J., Mckay, J., Maruff, P., & Anderson, V. (2006). The development of sustained attention in children: The effect of age and task load. *Child Neuropsychology*, *12*(3), 205-221. Retrieved from https://doi.org/10.1080/09297040500488522 (PMID: 16837396) doi: 10.1080/09297040500488522

Biocca, F. (1997). The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication*, *3*(2), JCMC324. (publisher=Oxford University Press Oxford, UK)

Blanck, P. D., Reis, H. T., & Jackson, L. (1984, Mar 01). The effects of verbal reinforcement of intrinsic motivation for sex-linked tasks. *Sex Roles*, *10*(5), 369–386. Retrieved from https://doi.org/10.1007/BF00287554 doi: 10.1007/BF00287554

Blatchford, P., & Russell, A. (2020). *Rethinking class size: The complex story of impact on teaching and learning*. UCL Press.

Blumenfeld, P. C., Kempler, T. M., & Krajcik, J. S. (2006). Motivation and cognitive engagement in learning environments [Book Section]. In R. K. Sawyer (Ed.), *The cambridge handbook of: The learning sciences.* (p. 475-488). New York, NY, US: Cambridge University Press.

Blut, M., Wang, C., Wünderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other ai. *Journal of the Academy of Marketing Science*, *49*(4), 632–658.

Boerma, T., Chiat, S., Leseman, P., Timmermeister, M., Wijnen, F., & Blom, E. (2015). A quasi-universal nonword repetition task as a diagnostic tool for bilingual children learning dutch as a second language. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1747-1760. doi: https://doi.org/10.1044/2015_JSLHR-L-15-0058

Boyer, K. E., Phillips, R., Wallis, M., Vouk, M., & Lester, J. (2008). Balancing cognitive and motivational scaffolding in tutorial dialogue. In *Intelligent tutoring systems* (Vol. 5091 LNCS, pp. 239–249). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-540-69132-7-28

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In IEEE (Ed.), *Ieee/rsj international conference on intelligent robots and systems* (p. 708-713). Conference on Intelligent Robots and Systems.

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology*, *68*, 627-652.

Brophy, J. (1981). Teacher praise: A functional analysis. *Review of Educational Research*, *51*(1), 5-32. Retrieved from https://doi.org/10.3102/00346543051001005 doi: 10.3102/00346543051001005

Byrne, D. (1987). *Techniques for classroom interaction*. Longman.

Cameron, J., & Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research*, *64*(3), 363-423. Retrieved from https://doi.org/10.3102/00346543064003363 doi: 10.3102/00346543064003363

Caporael, L. R. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior*, *2*(3), 215-234.

Casey, B. M., Andrews, N., Schindler, H., Kersh, J. E., Samper, A., & Copley, J. (2008). The development of spatial skills through interventions involving block building activities. *Cognition and Instruction*, *26*(3), 269–309.

&

Chandra, S., Paradeda, R., Yin, H., Dillenbourg, P., Prada, R., & Paiva, A. (2018). Do children perceive whether a robotic peer is learning or not? In (p. 41-49). ACM/IEEE International Conference on Human-Robot Interaction.

Chang, C.-W., Lee, J.-H., Chao, P.-Y., Wang, C.-Y., & Chen, G.-D. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Journal of Educational Technology & Society*, *13*(2), 13–24.

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).

Chen, H., Park, H. W., & Breazeal, C. (2020). Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers & Education*, *150*, 103836. Retrieved from https://www.sciencedirect.com/science/article/pii/S0360131520300373 doi: https://doi.org/10.1016/j.compedu.2020.103836

Chiat, S. (2015). Non-word repetition. In . N. M. E. S. Armon-Lotem J. de Jong (Ed.), *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*. New York, NY, US: Springer Science & Business Media. doi: 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. *Journal of developmental and Physical Disabilities*, *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*.

Coninx, A., Baxter, P., Oleari, E., Bellini, S., Bierman, B., Blanson Henkemans, O., ... Belpaeme, T. (2015). Towards Long-Term Social Child-Robot Interaction: Using Multi-Activity Switching to Engage Young Users. *Journal of Human-Robot Interaction*, *5*(1), 32. doi: 10.5898/jhri.5.1.coninx

Corpus, J. H., & Lepper, M. R. (2007). The effects of person versus performance praise on children's motivation: Gender and age as moderating factors. *Educational Psychology*, *27*(4), 487–508. Retrieved from https://doi.org/10.1080/01443410601159852 doi: 10.1080/01443410601159852

Dautenhahn, K. (2004, September). Robots we like to live with?! - a developmental perspective on a personalized, life-long robot companion. *RO-MAN th IEEE International Workshop on Robot and Human Interactive Communication*, *13*, 17-22.

Dautenhahn, K., & Werry, I. (2004). Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition*, *12*(1), 1-35. doi: https://doi.org/10.1075/pc.12.1.03dau

Davison, D. P., Wijnen, F. M., Charisi, V., van der Meij, J., Reidsma, d., & Evers, v. (2021). Words of encouragement: how praise delivered by a social robot changes children's mindset for learning. *Journal on multimodal user interfaces*, *15*(1), 61–76.

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, *19*(2), 109–134. doi: 10.1016/0092-6566(85)90023-6

Deci, E. L., & Ryan, R. M. (2000). Self-determination Theory and the Facilitation of Intrinsic Motivation, Social development, and Well-Being. *American Psychologist*, *55*(1), 68–78.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist*, *26*(3-4), 325-346. Retrieved from https://doi.org/10.1080/00461520.1991.9653137 doi: 10.1080/00461520.1991.9653137

de Haas, M., Aroyo, A. M., Barakova, E., Haselager, P., & Smeekens, I. (2016). The effect of a semi-autonomous robot on children. In *Proceedings of the 2016 ieee 8th international conference on intelligent systems (is)* (pp. 376–381).

de Haas, M., & Conijn, R. (2020). Carrot or stick: The effect of reward and punishment in robot assisted language learning. In *Companion of the 2020 acm/ieee international conference on human-robot interaction* (p. 177–179). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3371382.3378349 doi: 10.1145/3371382.3378349

Dempsey, J. V., & Sales, G. C. (1993). *Interactive instruction and feedback*. Englewood Cliffs, New Jersey, USA: Educational Technology Publications.

de Nooijer, J. A., van Gog, T., Paas, F., & Zwaan, R. A. (2013). Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. *Acta Psychologica*, *144*(1), 173-179.

&

de Wit, J., Brandse, A., Krahmer, E., & Vogt, P. (2020). Varied human-like gestures for social robots: Investigating the effects on children's engagement and language learning. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction* (p. 359–367). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3319502.3374815 doi: 10.1145/3319502.3374815

de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., ... Vogt, P. (2018). The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (pp. 50–58). Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3171221.3171277 doi: 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).

DiSalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002, June). All robots are not created equal: the design and perception of humanoid robot heads. In *Proceedings of the 4th conference on designing interactive systems: Processes, practices, methods, and techniques* (p. 321-326).

Dörnyei, Z. (1998). Motivation in second and foreign language learning. *Language Teaching*, *31*(3), 117–135. doi: 10.1017/S026144480001315X

Drijvers, L., Vaitonytė, J., & Özyürek, A. (2019). degree of language experience modulates visual attention to visible speech and iconic gestures during clear and degraded speech comprehension. *Cognitive Science*, *43*(10), e12789.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3-4), 177-190. doi: https://doi.org/10.1016/S0921-8890(02)00374-3

Dunn, L. M., Dunn, L. M., & Schlichting, L. (2005). *Peabody picture vocabulary test-iii-nl*. Amsterdam: Amsterdam: Pearson.

Dweck, C. S., & Bush, E. S. (1976). Sex differences in learned helplessness: I. differential debilitation with peer and adult evaluators. *developmental Psychology*, *12*(2), 147.

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145-157. Retrieved from https://www.sciencedirect.com/science/article/pii/S0959475211000806 doi: https://doi.org/10.1016/j.learninstruc.2011.10.001

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. doi: 10.1017/S0272263102002024

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143-155.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864.

Eyssel, F., Kuchenbrandt, D., Hegel, F., & de Ruiter, L. (2012, September). Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In (p. 851-857). doi: 10.1109/ROMAN.2012.6343858

Ferreira, M. J., Nisi, V., Melo, F., & Paiva, A. (2017). Learning and teaching biodiversity through a storyteller robot. In *International conference on interactive digital storytelling* (pp. 367–371).

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using r*. Sage publications.

Filippini, C., Spadolini, E., Cardone, D., Bianchi, D., Preziuso, M., Sciarretta, C., ... Merla, A. (2020). Facilitating the child–robot interaction by endowing the robot with the capability of understanding the child engagement: The case of mio amico robot. *International Journal of Social Robotics*, 1–13.

Fink, J. (2012, October). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In Springer (Ed.), *International conference on social robotics* (p. 199-208).

Flink, C., Boggiano, A. K., Main, D. S., Barrett, M., & Katz, P. A. (1992). Children's achievement-related behaviors: The role of extrinsic and intrinsic motivational orientations. *Achievement and motivation: A social-developmental perspective*, 189–214.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59-109. Retrieved from https://doi.org/10.3102/00346543074001059 doi: 10.3102/00346543074001059

Gaertner, B. M., Spinrad, T. L., & Eisenberg, N. (2008). Focused attention in toddlers: Measurement, stability, and relations to negative emotion and parenting. *Infant and Child development: An International Journal of Research and Practice*, *17*(4), 339–363.

&

Gordon, G., Spaulding, S., Kory-Westlund, J. M., Lee, J. J., Plummer, L., Martinez, M., ... Breazeal, C. (2016). Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the thirtieth aaai conference on artificial intelligence* (pp. 3951–3957). AAAI Press.

Hall, C. J. (2002). The automatic cognate form assumption: Evidence for the parasitic model of vocabulary development. *IRAL - International Review of Applied Linguistics in Language Teaching*, *40*(2), 69–87. doi: 10.1515/iral.2002.008

Halliday, S. E., Calkins, S. D., & Leerkes, E. M. (2018). Measuring preschool learning engagement in the laboratory. *Journal of Experimental Child Psychology*, *167*, 93-116. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022096517301029 doi: https://doi.org/10.1016/j.jecp.2017.10.006

Hamilton, V. J., & Gordon, D. A. (1978). Teacher-Child Interactions in Preschool and Task Persistence. *American Educational Research Journal*, *15*(3), 459–466. doi: 10.3102/00028312015003459

Hattie, J., & Gan, M. (2011). Instruction based on feedback. *Handbook of research on learning and instruction*, 249–271.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. Retrieved from http://rer.sagepub.com/cgi/doi/10.3102/003465430298487 doi: 10.3102/003465430298487

Heath, S., Durantin, G., Boden, M., Hensby, K., Taufatofua, J., Olsson, O., ... Wiles, J. (2017). spatiotemporal aspects of engagement during dialogic storytelling child–robot interaction. *Frontiers in Robotics and AI*, *4*, 27.

Hegel, F., Krach, S., Kircher, T., Wrede, B., & Sagerer, G. (2008, August). Understanding social robots: A user study on anthropomorphism. In *17th ieee international symposium on robot and human interactive communication* (p. 574-579). IEEE.

Henderlong, J., & Lepper, M. R. (2002). The effects of praise on children's intrinsic motivation: a review and synthesis. *Psychological bulletin*, *128*(5), 774–795. doi: 10.1037/0033-2909.128.5.774

Henry, S. E., Medway, F. J., & Scarbro, H. A. (1979). Sex and locus of control as determinants of children's responses to peer versus adult praise. *Journal of Educational Psychology*, *71*(5), 604.

Hindriks, K. V., & Liebens, S. (2019). A robot math tutor that gives feedback. In M. A. Salichs et al. (Eds.), *Social robotics* (pp. 601–610). Cham: Springer International Publishing.

Hockema, S. A., & Smith, L. B. (2009). Learning your language, outside-in and inside-out. *Linguistics*, *47*(2), 453-479. doi: https://doi.org/10.1515/LING.2009.016

Hoff, E. (2013). Interpreting the early language trajectories of children from low-ses and language minority homes: implications for closing achievement gaps. *developmental psychology*, *49*(1), 4.

Hood, d., Lemaignan, S., & Dillenbourg, P. (2015). When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 83–90).

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Sage.

Inoue, K., Lala, D., Takanashi, K., & Kawahara, T. (2018). Engagement recognition by a latent character model based on multimodal listener behaviors in spoken dialogue. *APSIPA Transactions on Signal and Information Processing*, *7*, e9. doi: 10.1017/ATSIP.2018.11

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).

Ishii, R., Ooko, R., Nakano, Y. I., & Nishida, T. (2013). Effectiveness of gaze-based engagement estimation in conversational agents. In Y. I. Nakano, C. Conati, & T. Bader (Eds.), *Eye gaze in intelligent user interfaces: Gaze-based analyses, models and applications* (pp. 85–110). London: Springer London. Retrieved from https://doi.org/10.1007/978-1-4471-4784-8_6 doi: 10.1007/978-1-4471-4784-8_6

Iverson, J. M. (2010). developing language in a developing body: The relationship between motor development and language development. *Journal of Child Language*, *37*(2), 229-261.

Jacq, A., Lemaignan, S., Garcia, F., Dillenbourg, P., & Paiva, A. (2016, March). 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). doi: 10.1109/HRI.2016.7451758

Jang, H. (2008). Supporting students' motivation, engagement, and learning during an uninteresting activity. *Journal of Educational Psychology*, *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).

&

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. *Frontiers in Robotics and 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 *Proceedings of the 2018 chi conference on human factors in computing systems* (p. 1–13). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3173574.3174069 doi: 10.1145/3173574.3174069

Jipson, J. L., & Gelman, S. A. (2007). Robots and rodents: Children's inferences about living and nonliving kinds. *Child development*, *78*(6), 1675-1688.

Kahn Jr, P. H., Gary, H. E., & Shen, S. (2013). Children's social relationships with current and near-future robots. *Child development Perspectives*, *7*(1), 32-37.

Kahn Jr, P. H., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., ... Shen, S. (2012). "robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *developmental Psychology*, *48*(2), 303.

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction*, *19*(1-2), 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. doi: 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 development Perspectives*, *12*(3), 146–151. doi: 10.1111/cdep.12277

Kelly, S. D., Mcdevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, *24*(2), 313-334.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction* (pp. 67–74). New York, NY, USA: ACM.

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2016). Social robot tutoring for child second language learning. In *The eleventh acm/ieee international conference on human robot interaction* (p. 231-238). Christchurch, New Zealand: IEEE Press. doi: 10.1109/HRI.2016.7451757

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., ... Belpaeme, T. (2017). Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations. In *Acm/ieee international conference on human-robot interaction* (Vol. Part F1271, pp. 82–90). doi: 10.1145/2909824.3020229

Kidd, C. D. (2003). *Sociable robots: The role of presence and task in human-robot interaction (doctoral dissertation*. Massachusetts Institute of Technology.

Kim, Y., Butail, S., Tscholl, M., Liu, L., & Wang, Y. (2020). An exploratory approach to measuring collaborative engagement in child robot interaction. In *Proceedings of the tenth international conference on learning analytics and knowledge* (p. 209–217). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3375462.3375522 doi: 10.1145/3375462.3375522

Kluger, A. N., & deNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. doi: 10.1037/0033-2909.119.2.254

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).

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

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. *International Journal of Social Robotics*, 1–17.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. doi: 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 *Acm/ieee international conference on human-robot interaction* (Vol. 02-05-Marc, pp. 65–66). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2701973.2701989 doi: 10.1145/2701973.2701989

&

Kory-Westlund, J. M., & Breazeal, C. (2019a). Assessing children's perceptions and acceptance of a social robot. In *Proceedings of the 18th acm international conference on interaction design and children* (p. 38-50). 18th ACM International Conference on Interaction design and Children.

Kory-Westlund, J. M., & Breazeal, C. (2019b). A long-term study of young children's rapport, social emulation, and language learning with a peer-like robot playmate in preschool. *Frontiers in Robotics and AI*, *6*, 81. Retrieved from https://www.frontiersin.org/article/10.3389/frobt.2019.00081 doi: 10.3389/frobt.2019.00081

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* (p. 28-29).

Kory-Westlund, J. M., Dickens, L., Jeong, S., Harris, P. L., deSteno, D., & Breazeal, C. L. (2017). Children use non-verbal cues to learn new words from robots as well as people. *International Journal of Child-Computer Interaction*, *13*, 1-9. Retrieved from https://www.sciencedirect.com/science/article/pii/S2212868916300538 doi: https://doi.org/10.1016/j.ijcci.2017.04.001

Kory-Westlund, J. M., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., ... Breazeal, C. L. (2017). Flat vs. expressive storytelling: young children's learning and retention of a social robot's narrative. *Frontiers in human neuroscience*, *11*, 295. Retrieved from https://www.frontiersin.org/article/10.3389/fnhum.2017.00295 doi: 10.3389/fnhum.2017.00295

Kory-Westlund, J. M., Martinez, M., Archie, M., Das, M., & Breazeal, C. (2016). Effects of framing a robot as a social agent or as a machine on children's social behavior. In *2016 25th ieee international symposium on robot and human interactive communication (ro-man)* (p. 688-693). IEEE.

Laevers, F. (2003). Experiential education: Making care and education more effective through well-being and involvement. *Involvement of children and teacher style. Insights from an international study on experiential education*, 13–24.

Laevers, F. (2005). Well-being and involvement in care settings. a process-oriented self-evaluation instrument (sic's). *Kind en gezin; Research Centre for Experiential Education; Leuven*, 1-20.

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

Laevers, F., Daems, M., De Bruyckere, G., Declercq, B., Moons, J., Silkens, K., ... van Kessel, M. (2005). *Ziko. zelfevaluatie-instrument voor welbevinden en betrokkenheid van kinderen in de opvang.* Brussel: Kind & Gezin.

Leeuwestein, H., Barking, M., Sodacı, H., Oudgenoeg-Paz, O., Verhagen, J., Vogt, P., ... de Haas, M.,... Leseman, P. (2021). Teaching turkish-dutch kindergartners dutch vocabulary with a social robot: Does the robot's use of turkish translations benefit children's dutch vocabulary learning? *Journal of Computer Assisted Learning*, *37*(3), 603–620.

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2014). Empathic robots for long-term interaction. *International Journal of Social 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).

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

Leite, I., McCoy, M., Lohani, M., Ullman, D., Salomons, N., Stokes, C., ... Scassellati, B. (2017). Narratives with robots: The impact of interaction context and individual differences on story recall and emotional understanding. *Frontiers in Robotics and AI*, *4*, 29. Retrieved from https://www.frontiersin.org/article/10.3389/frobt.2017.00029 doi: 10.3389/frobt.2017.00029

Lemaignan, S., Fink, J., & Dillenbourg, P. (2014, March). The dynamics of anthropomorphism in robotics. In *In 9th acm/ieee international conference on human-robot interaction* (p. 226-227).

Lemaignan, S., Fink, J., Dillenbourg, P., & Braboszcz, C. (2014). *The cognitive correlates of anthropomorphism*. In 2014 Human-Robot Interaction Conference, Workshop "HRI: a bridge between Robotics and Neuroscience".

Lemaignan, S., Fink, J., Mondada, F., & Dillenbourg, P. (2015). You're doing it wrong! studying unexpected behaviors in child-robot interaction. In Springer (Ed.), *International conference on social robotics* (p. 390-400). Conference on Social Robotics.

Leseman, P., Henrichs, L., Blom, W., & Verhagen, J. (2019). Young mono-and bilingual children's exposure to academic language as related to language development and school achievement. *Learning through language*, 205.

Leyzberg, D., Spaulding, S., & Scassellati, B. (2014). Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 acm/ieee international conference on human-robot interaction* (pp. 423–430). New York, NY, USA: ACM.

Ligthart, M., Neerincx, M. A., & Hindriks, K. V. (2019). Getting acquainted for a long-term child-robot interaction. In M. A. Salichs et al. (Eds.), *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 *Proceedings of the 2020 acm/ieee international conference on human-robot interaction* (p. 409–418). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3319502.3374826 doi: 10.1145/3319502.3374826

Linnenbrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs instudent engagement and learning intheclassroom. *Reading & Writing Quarterly*, *19*(2), 119-137. Retrieved from https://doi.org/10.1080/10573560308223 doi: 10.1080/10573560308223

Long, M. H. (2006). *Problems in SLA (Second Language Acquisition Research)* (ERIC, Ed.). 10 Industrial Avenue, Mahwah: Lawrence Erlbaum Associates (Bks). Retrieved from http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20{&}path=ASIN/0805860843

Lyster, R., & Ranta, L. (1997). Corrective Feedback and Learner Uptake: Negotiation of Form in Communicative Classrooms. *Studies in Second Language Acquisition*, *19*(1), 37–66. doi: 10.1017/s0272263197001034

Lyster, R., & Saito, K. (2010). Oral feedback in classroom sla. *Studies in Second Language Acquisition*, *32*(02), 265–302.

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

Mackey, A., & Oliver, R. (2002). Interactional feedback and children's l2 development. *System*, *30*(4), 459-477. Retrieved from https://www.sciencedirect.com/science/article/pii/S0346251X02000490 doi: https://doi.org/10.1016/S0346-251X(02)00049-0

Mackey, A., Oliver, R., & Leeman, J. (2003). Interactional input and the incorporation of feedback: An exploration of NS–NNS and NNS–NNS adult and child dyads. *Language Learning*, *53*(1), 35–66.

Mackey, A., & Silver, R. E. (2005). Interactional tasks and English L2 learning by immigrant children in Singapore. *System*, *33*(2), 239–260. doi: 10.1016/j.system.2005.01.005

Mair, P., & Wilcox, R. (2020). Robust statistical methods in r using the wrs2 package. *Behavior research methods*, *52*(2), 464–488.

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research*, *80*(3), 300-335.

Masgoret, A. M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: a meta–analysis of studies conducted by Gardner and associates. *Language Learning*, *53*(1), 123–163.

Mashburn, A. J., Justice, L. M., Downer, J. T., & Pianta, R. C. (2009). Peer effects on children's language achievement during pre-kindergarten. *Child development*, *80*(3), 686–702.

Mazzoni, E., & Benvenuti, M. (2015). A Robot-partner for Preschool Children learning English using Socio-Cognitive Conflict. *Educational Technology and Society*, *18*(4), 474–485.

Michaelis, J. E., & Mutlu, B. (2018). Reading socially: Transforming the in-home reading experience with a learning-companion robot. *Science Robotics*, *3*, 21.

Midden, C., & Ham, J. (2014). The power of negative feedback from an artificial agent to promote energy saving behavior. In *International conference of design, user experience, and usability* (pp. 328–338).

Monaco, C., Mich, O., Ceol, T., & Potrich, A. (2018). *Investigating mental representations about robots in preschool children. arxiv.* (preprint)

Moon, A., Troniak, D. M., Gleeson, B., Pan, M. K., Zheng, M., Blumer, B. A., ... Croft, E. A. (2014). Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proceedings of the 2014 acm/ieee international conference on human-robot interaction* (p. 334-341).

Morgan, G. A., Harmon, R. J., & Maslin-Cole, C. A. (1990). Mastery motivation: definition and measurement. *Early Education and development*, *1*(5), 318-339. Retrieved from https://doi.org/10.1207/s15566935eed0105_1 doi: 10.1207/s15566935eed0105_1

Morris, B. J., & Zentall, S. R. (2014). High fives motivate: The effects of gestural and ambiguous verbal praise on motivation. *Frontiers in Psychology*, *5*(AUG), 928. doi: 10.3389/fpsyg.2014.00928

&

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

Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., & Dong, J.-J. (2013). A review of the applicability of robots in education. *Journal of Technology in Education and Learning*, *1*(209-0015), 13.

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. *Frontiers in Psychology*, *5*, 733.

Mwangi, E., Barakova, E., Boladeras, M., Català, A., & Rauterberg, M. (2018, 08). Dyadic gaze patterns during child-robot collaborative gameplay in a tutoring interaction. In *27th ieee international symposium on robot and human interactive communication (ro-man)* (p. 856-86). doi: 10.1109/ROMAN.2018.8525799

Nakano, Y. I., & Ishii, R. (2010). Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on intelligent user interfaces* (pp. 139–148). New York, NY, USA: ACM.

Neto, I., Nicolau, H., & Paiva, A. (2021). Community based robot design for classrooms with mixed visual abilities children. In *Proceedings of the 2021 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3411764.3445135

Nijssen, S. R., Müller, B. C., Bosse, T., & Paulus, M. (2021). You, robot? the role of anthropomorphic emotion attributions in children's sharing with a robot. *International Journal of Child-Computer Interaction*, *30*, 100319. Retrieved from https://www.sciencedirect.com/science/article/pii/S2212868921000465 doi: https://doi.org/10.1016/j.ijcci.2021.100319

Obaid, M., Barendregt, W., Alves-Oliveira, P., Paiva, A., & Fjeld, M. (2015). designing robotic teaching assistants: interaction design students' and children's views. In *International conference on social robotics* (p. 502-511).

Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., & Peters, C. (2020). Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI*, *7*, 92. Retrieved from https://www.frontiersin.org/article/10.3389/frobt.2020.00092 doi: 10.3389/frobt.2020.00092

Okita, S. Y., & Schwartz, D. L. (2013). Learning by teaching human pupils and teachable agents: The importance of recursive feedback. *Journal of the Learning Sciences*, *22*(3), 375–412.

Oliver, R. (2000). Age differences in negotiation and feedback in classroom and pairwork. *Language Learning*, *50*(1), 119–151.

Oudgenoeg-Paz, O., Leseman, P. M., & Volman, M. (2015). Exploration as a mediator of the relation between the attainment of motor milestones and the development of spatial cognition and spatial language. *developmental Psychology*, *51*(9), 1241-1253.

Oxford, R., & Shearin, J. (1994). Language learning motivation: Expanding the theoretical framework. *The Modern Language Journal*, *78*(1), 12–28.

Park, E., Kim, K. J., & del Pobil, A. P. (2011). The effects of a robot instructor's positive vs. negative feedbacks on attraction and acceptance towards the robot in classroom. In B. Mutlu, C. Bartneck, J. Ham, v. Evers, & T. Kanda (Eds.), *Social robotics: Third international conference, icsr 2011, amsterdam, the netherlands, november 24-25, 2011. proceedings* (pp. 135–141). Berlin, Heidelberg: Springer Berlin Heidelberg.

Park, H. W., Gelsomini, M., Lee, J. J., & Breazeal, C. (2017). Telling Stories to Robots: The Effect of Backchanneling on a Child's Storytelling. In *Acm/ieee international conference on human-robot interaction* (Vol. Part F1271, pp. 100–108). doi: 10.1145/2909824.3020245

Pereira, A., Martinho, C., Leite, I., & Paiva, A. (2008). icat, the chess player: the influence of embodiment in the enjoyment of a game. In *Proceedings of the 7th international joint conference on autonomous agents and multiagent systems-volume 3* (p. 1253-1256).

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*, 1-1. doi: 10.1109/TAFFC.2020.2980275

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018, March). What is human-like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (p. 105-113).

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

Pot, E., Monceaux, J., Gelin, R., & Maisonnier, B. (2009, September). Choregraphe: A graphical tool for humanoid robot programming. In *Ro-man 2009-the 18th ieee international symposium on robot and human interactive communication* (p. 46-51).

&

Reidsma, d., Charisi, V., Davison, D., Wijnen, F., van der Meij, J., Evers, v., ... Verschure, P. (2016). The easel project: Towards educational human-robot symbiotic interaction. In N. F. Lepora, A. Mura, M. Mangan, P. F. Verschure, M. desmulliez, & T. J. Prescott (Eds.), *Biomimetic and biohybrid systems* (pp. 297–306). Cham: Springer International Publishing.

Resing, W. C., Bakker, M., Elliott, J. G., & Vogelaar, B. (2019). Dynamic testing: Can a robot as tutor be of help in assessing children's potential for learning? *Journal of Computer Assisted Learning*, *35*(4), 540–554. doi: 10.1111/jcal.12358

Rich, C., Ponsleur, B., Holroyd, A., & Sidner, C. L. (2010). Recognizing engagement in human-robot interaction. In *Proceedings of the 5th acm/ieee international conference on human-robot interaction* (p. 375–382). IEEE Press.

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009, March). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th acm/ieee international conference on human robot interaction* (p. 245-246).

Rintjema, E., van den Berghe, R., Kessels, A., de Wit, J., & Vogt, P. (2018, 3). A robot teaching young children a second language: The effect of multiple interactions on engagement and performance. In *Companion of the 2018 acm/ieee international conference on human-robot interaction* (pp. 219–220). ACM.

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research*, *81*(4), 493-529. Retrieved from https://doi.org/10.3102/00346543114 21793 doi: 10.3102/0034654311421793

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.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, *55*(1), 68.

Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010). Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1613–1622).

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, *5*(3), 313-323. (Springer)

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).

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).

Schodde, T., Hoffmann, L., & Kopp, S. (2017). How to manage affective state in child-robot tutoring interactions? In *2017 international conference on companion technology (icct)* (pp. 1–6).

Schodde, T., Hoffmann, L., Stange, S., & Kopp, S. (2019, December). Adapt, explain, engage—a study on how social robots can scaffold second-language learning of children. *J. Hum.-Robot Interact.*, *9*(1). Retrieved from https://doi.org/10.1145/3366422 doi: 10.1145/3366422

Schrooten, W., & Vermeer, A. (1994). *Woorden in het basisonderwijs. 1500 woorden aangeboden aan leerlingen. [Words in primary education. 1500 words presented to students].* Tilburg University Press.

Sciutti, A., Rea, F., & Sandini, G. (2014). When you are young,(robot's) looks matter. developmental changes in the desired properties of a robot friend. In *In 2014 ro-man: The 23rd ieee international symposium on robot and human interactive communication* (p. 567-573).

Scott, V. M., & de la Fuente, M. J. (2008). What's the problem? L2 learners' use of the L1 during consciousness-raising, form-focused tasks. *Modern Language Journal*, *92*(1), 100–113. doi: 10.1111/j.1540-4781.2008.00689.x

Serholt, S., & Barendregt, W. (2016). Robots Tutoring Children: Longitudinal Evaluation of Social Engagement in Child-Robot Interaction. In *Proceedings of the 9th nordic conference on human-computer interaction* (p. 64). New York, NY, USA: ACM. doi: 10.1145/2971485.2971536

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. doi: 10.3102/0034654307313795

Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). Where to look: A study of human-robot engagement. In *Proceedings of the 9th international conference on intelligent user interfaces* (p. 78–84). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/964442.964458 doi: 10.1145/964442.964458

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. Retrieved from http://www.sciencedirect.com/science/article/pii/S0004370205000512 doi: https://doi.org/10.1016/j.artint.2005.03.005

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632

Smakman, M., Vogt, P., & Konijn, E. A. (2021). Moral considerations on social robots in education: A multi-stakeholder perspective. *Computers & Education*, *174*, 104317. Retrieved from https://www.sciencedirect.com/science/article/pii/S0360131521001949 doi: https://doi.org/10.1016/j.compedu.2021.104317

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. doi: 10.1016/j.cognition.2007.06.010

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. *International Journal of Play*, *8*(1), 65-78. Retrieved from https://doi.org/10.1080/21594937.2019.1580338 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.

Tanaka, F., & Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, *1*(1).

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

Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., ... David, D. (2020). *Social engagement of children with autism during interaction with a robot.*

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *8*(2), 219-235.

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology*, *25*(6), 631-645.

Trowler, V. (2010). Student engagement literature review. *The higher education academy*, *11*(1), 1–15.

Tung, F.-W. (2011, July). Influence of gender and age on the attitudes of children towards humanoid robots. In *International conference on human-computer interaction* (p. 637-646). Springer.

Tung, F.-W. (2016). Child perception of humanoid robot appearance and behavior. *International Journal of Human-Computer Interaction*, *32*(6), 493-502.

van Minkelen, P., Gruson, C., van Hees, P., Willems, M., de Wit, J., Aarts, R., ... Vogt, P. (2020). Using self-determination theory in social robots to increase motivation in l2word learning. In *Acm/ieee international conference on human-robot interaction.*

van den Berghe, R., Oudgenoeg-Paz, O., Verhagen, J., Brouwer, S., de Haas, M., de Wit, J., ... Leseman, P. (2021). Individual differences in children's (language) learning skills moderate effects of robot-assisted second language learning. *Frontiers in Robotics and AI*, *8*, 259. Retrieved from https://www.frontiersin.org/article/10.3389/frobt.2021.676248 doi: 10.3389/frobt.2021.676248

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., & Leseman, P. (2019). Social robots for language learning: A review. *Review of Educational Research*, *89*(2), 259-295. Retrieved from https://doi.org/10.3102/0034654318821286 doi: 10.3102/0034654318821286

van Straten, C. L., Peter, J., & Kühne, R. (2020). Child–robot relationship formation: A narrative review of empirical research. *International Journal of Social Robotics*, *12*(2), 325–344. (Springer)

van Straten, C. L., Peter, J., Kühne, R., & Barco, A. (2020). Transparency about a robot's lack of human psychological capacities: Effects on child-robot perception and relationship formation. *ACM Transactions on Human-Robot Interaction (THRI)*, *9*(2), 1-22. (ACM New York, NY, USA)

Vázquez, M., Steinfeld, A., Hudson, S. E., & Forlizzi, J. (2014, June). 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.

Verhagen, J., Berghe, R. v. d., Oudgenoeg-Paz, O., Küntay, A., & Leseman, P. (2019). Children's reliance on the non-verbal cues of a robot versus a human. *Plos one*, *14*(12), e0217833. (Public Library of Science San Francisco, CA USA)

Vogt, P., de Haas, M., de Jong, C., Baxter, P., & Krahmer, E. (2017). Child-robot interactions for second language tutoring to preschool children. *Frontiers in Human Neuroscience*, *11*. doi: 10.3389/fnhum.2017.00073

Vogt, P., van den Berghe, R., de Haas, M., Hoffman, L., Kanero, J., Mamus, E., ... others (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). doi: 10.1109/HRI.2019.8673077

Vygotsky, L. (1978). Interaction between learning and development. In V. J.-S. Cole, S. Scriber, & E. Souberman (Eds.), *M* (p. 79-97). Cambridge: Harvard University Press.

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410.

Wellsby, M., & Pexman, P. M. (2014). developing embodied cognition: Insights from children's concepts and language processing. *Frontiers in Psychology*, *5*, 506.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of lrec* (Vol. 5, p. 1556-1559). Genoa, Italy.

Wojitas, O. (1998). Feedback? No, just give us the answers. *Times Higher Education Supplement*, *25*(September 1998), 109162. Retrieved from http://www.sythe.org/archive/index.php?t-688832.html

Woumans, E., Surmont, J., Struys, E., & Duyck, W. (2016). The longitudinal effect of bilingual immersion schooling on cognitive control and intelligence*. *Language Learning*, *66*(S2), 76-91. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12171 doi: https://doi.org/10.1111/lang.12171

Xu, T., Zhang, H., & Yu, C. (2016). See you see me: the role of eye contact in multimodal human-robot interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *6*(1), 1–22.

Zaga, C., de Vries, R. A., Li, J., Truong, K. P., & Evers, v. (2017, May). A simple nod of the head: The effect of minimal robot movements on children's perception of a low-anthropomorphic robot. In *Proceedings of the 2017 chi conference on human factors in computing systems* (p. 336-341).

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. doi: 10.1007/978-3-319-25554-5_70

Zaga, C., Truong, K. P., Lohse, M., & Evers, v. (2014, June). 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).

Zentall, S. R., & Morris, B. J. (2010). "Good job, you're so smart": The effects of inconsistency of praise type on young children's motivation. *Journal of Experimental Child Psychology*, *107*(2), 155–163. doi: 10.1016/j.jecp.2010.04.015

Zentall, S. R., & Morris, B. J. (2012). A critical eye: Praise directed toward traits increases children's eye fixations on errors and decreases motivation. *Psychonomic Bulletin and Review*, *19*(6), 1073–1077. doi: 10.3758/s13423-012-0294-y

Zhang, P., & de Haas, M. (2020). Effects of pitch gestures on learning chinese orthography with a social robot. In *Companion of the 2020 acm/ieee international conference on human-robot interaction* (p. 180–182). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3371382.3378350 doi: 10.1145/3371382.3378350

&

Appendix

This appendix provides the English and Dutch annotation guidelines for engagement coding as used in Chapters 2, 3 and 4.

English version coding scheme engagement

This manual is based on an extensively tested measuring instrument called ”ziko” (Laevers et al., 2005). Before you can get started, you need to get to know 1 of the concepts behind the instrument: engagement. It is important that you learn in advance to look specifically at children and know how to work with the instrument. You can only enter the scores correctly if you have mastered the manual. The preparation of the self-evaluation is of great importance. If you want more information and help with practicing ziko, go to ecego. You can find more information on the website of child and family (`www.kindengezin.be`) and of ecego (`www.cego.be`) or this publication:
Laevers, F. (2005). Well-being and involvement in care settings. a process-oriented self-evaluation instrument (sic's). *Kind en gezin; Research Centre for Experiential Education; Leuven*, 1-20.

What is engagement?

A child who is engaged is, in a way, "completely absorbed" in his activity: Playing with blocks, modelling clay or puzzling, listening to a story, talking to others, it is a very specific experience that you can recognize in both babies and adults. Engagement is something very special. Anyone will get surprised by it, just by looking at children. You feel intuitively that you must not disturb the game. If there is engagement, we know that children are addressing their possibilities and that they are 'developing': they learn at a deeper level, and they become more competent. Engagement includes:

- **Motivation**. When you are engaged, you feel attracted by the activity, so you really are interested. You do not get engagement if others ask you or oblige you to do things. Your motivation arises from yourself, so although this may have been assigned to you, you are actively working on it yourself.

- **Intense mental activity**. When you are engaged you find yourself completely open to experiences: the impressions you gain are intense. Body sensations and movement experiences, colours and sounds, smells and flavours have a hue and depth that would otherwise remain unnoticed. You use your imagination and cognitive ability fully. In the absent of engagement, the sensations are not full lived through, that is, they are superficial.
- **Satisfaction**. Engagement is a wonderful condition: you are ecstatic. What you experience is energy that passes through your body. Children spontaneously take initiatives that will keep them in that state. Playing is an excellent manner in which they find this satisfaction. If engagement is lacking, you get bored, a feeling of emptiness and frustration.
- **Exploration urge**. The source for engagement is the urge to explore, the urge to go around the world to gain sensory impressions, to get a grip on reality. Initially, that 'getting a hold' can be taken literally: touching and grasping whatever comes close. Gradually it is more about "understanding" reality.
- **At the limit of your capabilities**. Engagement is possible if an activity is a challenge, neither too easy nor too difficult. When engaged, people move at the limit of their capabilities. They use their abilities to the fullest, they give the best of themselves - whether we are talking about babies or adults, or about children with poor mental development or about highly intelligent people.

YOUR JOB You will determine the engagement of the child. You will observe the child for two minutes. Give each child a score for engagement based on a five-point scale (1-5), where 1 is low engaged and 5 is highly engaged. You can also give half points, so the child can also be 3.5 engaged. When watching the video, remember that it is a snapshot, so it is possible that the same child scores low on engagement in one moment a higher in another moment. That means that if the child shows a higher engagement in the beginning of the clip compared to the last part; then you mediate the score of engagement over these two values. This mediating also depends on the period of time that the child shows this level of engagement, so for instance if the child shows a third of the video clip a high engagement (5) and shows a lower engagement (3) during two-thirds of the clip. The final level for engagement will be between 3.5 and 4. Therefore, it is useful to make notes on how engaged the child is and why you think so. We measure two types of engagement: task engagement and robot engagement.

&

Measuring task engagement

Task engagement defines how the child is engaged with the task. A child can be engaged with a tablet, blocks in front of him/her, but also with the robot when the robot asks the child to do something (such as repeat and copy). If the child looks at the robot because the robot is talking, the child is still engaged in the task. Also, if the child looks at the robot when the robot shows a gesture, child task engagement remains. After all, repeating and gesturing are part of the task. Only in the case that the child focuses on something else during the task or looks at the robot for no reason do you score a lower level for the child task engagement. This also means that you do not measure how committed the child is with the robot, since that is the focus of the other engagement scale. Task engagement is also accompanied by errors in the game; in general, a lower engagement leads to more errors by the child. But, as you probably recognize by conducting the experiments yourself, the system sometimes saw errors that were actually not wrong. In this case it is up to you not to include these errors in your child engagement score. But always be aware that a low score on learning is not equal to a low engagement score, it is only possible that the two are related.

Measuring robot engagement

Robot engagement only looks at how the child is engaged with the robot. This is not related to the task. The child can be engaged with the robot without the child performing the task. Child-robot engagement is determined by how often the child talks to the robot and looks at the robot. Only repeating a target word is not a sign of child-robot engagement, since the children in the tablet condition also talk after the tablet. If the child also looks at the robot when repeating the target word, it does count as child-robot engagement. Children who imitate the gestures of the robot also show a high engagement. A child who only looks at the tablet and ignores the robot (tries to ignore it) will score lower.

Table 1: Scale for task engagement

Level	Engagement	Examples
1	Very low	*The child shows virtually no activity:* - No concentration: staring, dreaming away; - An absent, passive attitude; - No targeted activity, aimless actions, not triggering anything; - Meaningless ticking on the screen in order to continue; - Only concerned with the experiment leader and not with the task; - No signs of exploration and interest; - Do not absorb anything, no mental activity
2	Low	*The child shows some activity, but is regularly interrupted:* - Limited concentration: looking away, fidgeting, dreaming; - Easily distracted; - Tasks are performed to a limited extent
3	Mediocre	*There is activity all the time, but not really concentrated* - The child is routine, fleeting; - Has limited motivation, does not feel challenged, shows no real commitment; - Does not gain in-depth experience; - Is not absorbed by what it does; - Only uses his capacities in moderation; - The activity does not touch the imagination and the mind of the child. - Most tasks are performed.
4	High	*There are usually signs of engagement:* - The child is totally absorbed in his game; - There is usually concentration, but sometimes the attention drops; - The child feels challenged, there is a certain drive; - Uses its abilities; - Appeals to the imagination and the mind.
5	Very high	*The child is continuously busy and becomes absorbed in his activity:* - Is continuously concentrated, absorbed by the activity, forgets about the time; - Is very motivated, feels strongly addressed; - Cannot be distracted; - Looks carefully at the task, pays attention to details; - Is constantly appealing to all its capacities and possibilities; - There is a strong mental activity: the imagination and the mind run at full speed; - Gains profound new experiences; - Enjoy being so passionately engaged.

Table 2: Scale for robot engagement

Level	Engagement	Examples
1	Very low	*The child shows virtually no interaction with the robot:* - Ignores the robot completely; - Has a closed (body) position towards the robot; - An absent, passive attitude; - No targeted activity, aimless actions, not triggering anything; - No signs of interest in the robot.
2	Low	*The child shows some interaction with the robot, but this is regularly interrupted:* - Limited looking at the robot; - Easily distracted from the robot.
3	Mediocre	*There is activity all the time with the robot* - The child works routinely, being fleeting; - Has limited motivation, does not feel challenged, shows no real commitment; - Has an open (body) attitude towards the robot; - Does not gain in-depth experience; - Is not absorbed by the activity; - Aimlessly touching the robot.
4	High	*There are usually signs of robot engagement:* - The child is totally absorbed in his game with the robot; - There is usually joint attention; - There is usually concentration, but sometimes the attention drops; - The child feels challenged, there is a certain drive.
5	Very high	*The child is continuously absorbed in his activity with the robot:* - Is continuously focused on the robot; - Feels strongly addressed; - Cannot be distracted from the robot; - Looks carefully at the robot, pays attention to details; - Talks to the robot; - Copies gestures; (in the iconic gesture condition); - There is joint attention - Enjoy being so passionately engaged with the robot.

Nederlandse versie codeerschema engagement

Deze handleiding is gebaseerd op een uitgebreid getest meetinstrument genoemd "ziko". Vo-ordat je aan de slag kan, moet je 1 van de begrippen achter het instrument leren kennen: engagement. Het is belangrijk dat je vooraf leert om gericht te kijken naar kinderen en weet hoe je moet werken met het instrument. Enkel als je de handleiding onder de knie hebt, kan je de scores juist invullen. De voorbereiding van de zelfevaluatie is van groot belang. Wil je meer informatie en hulp bij het inoefenen van ziko? Daarvoor kan je terecht bij ecego. Je vindt meer informatie op de website van kind en gezin (`www.kindengezin.be`) en van ecego (`www.cego.be`) of in deze publicatie:
Laevers, F., Daems, M., De Bruyckere, G., Declercq, B., Moons, J., Silkens, K., ... van Kessel, M. (2005). *Ziko. zelfevaluatie-instrument voor welbevinden en betrokkenheid van kinderen in de opvang.* Brussel: Kind & Gezin.

Wat is engagement

Een kind dat engaged is, wordt als het ware 'helemaal opgeslorpt' in zijn activiteit: Spelen met blokken, boetseren of puzzelen, luisteren naar een verhaal, met anderen praten, het is een heel aparte beleving die je zowel bij baby's als bij volwassenen kan herkennen. Engagement is iets heel bijzonders. Iedereen die gewoon naar kinderen kijkt, wordt erdoor verrast. Je voelt intuïtief aan dat je het spel niet mag verstoren. Is er engagement, dan weten we dat kinderen hun mogelijkheden aanspreken en dat ze 'in ontwikkeling' zijn: ze leren op een dieper niveau, ze worden echt competenter. Engagement bestaat uit de volgende elementen:

- **Motivatie** Als je engaged bent, voel je je aangesproken door de activiteit, dus ben je werkelijk geïnteresseerd. Engagement krijg je niet als je dingen alleen maar doet omdat anderen het vragen of er jou toe verplichten. Je motivatie komt vanuit jezelf, dit kan dus wel opgedragen zijn vanuit anderen, maar je bent er zelf actief mee bezig.
- **Intense mentale activiteit** Bij engagement stel je je helemaal open voor ervaringen: de indrukken die je opdoet zijn heel sterk. Lichaamsgewaarwordingen en bewegingservaringen, kleuren en klanken, geuren en smaken hebben een schakering en een diepte die er anders niet zijn. Je spreekt je verbeelding en je denkvermogen ten volle aan. Bij niet-betrokken activiteit zijn de gewaarwordingen niet doorleefd, dus oppervlakkig.
- **Voldoening** Engagement is een heerlijke toestand: je bent in vervoering. Wat je beleeft is energie die door je stroomt. Kinderen nemen spontaan steeds opnieuw initiatieven die hen

&

in die toestand brengen. Spel is de plek bij uitstek waarin ze deze genoegdoening vinden. Ontbreekt engagement, dan krijg je verveling, een gevoel van leegte en frustratie.

- **Exploratiedrang** De bron voor engagement is de ontdekkings- of exploratiedrang, de drang om de wereld te ervaren, om zintuiglijke indrukken op te doen, om greep te krijgen op de werkelijkheid. Aanvankelijk is dat 'greep krijgen' letterlijk te nemen: aanraken en grijpen wat in de buurt komt. Gaandeweg gaat het meer om het 'begrijpen' van de werkelijkheid.
- **Aan de grens van je mogelijkheden** Engagement is mogelijk als een activiteit een uitdaging is, niet te makkelijk en ook niet te moeilijk. Bij engagement bewegen mensen zich dus aan de grens van hun mogelijkheden. Ze spreken hun vermogens ten volle aan, ze geven het beste van zichzelf - of we het nu over baby's hebben of volwassenen, over kinderen met een zwakke mentale ontwikkeling of over hoogbegaafden.

Jouw taak Je gaat de engagement van het kind bepalen. Je observeert het kind gedurende een tweetal minuten. Geef elk kind een score voor engagement op basis van een vijfpuntenschaal. Je mag ook halve punten geven, dus het kind kan ook 3.5 engaged zijn. Bij het scannen gaat het om een momentopname, het kan dus zijn dat hetzelfde kind het ene fragment een lage engagement scoort en het andere moment een hogere engagement. Daarnaast kijk je naar de engagement over het gehele fragment. Laat het kind dus in het begin van het fragment een hogere engagement zien dan in het laatste gedeelte; dan middel je over deze twee waardes. Dit middelen laat je ook afhangen van de periode dat het kind deze engagement laat zien, als het kind dus een derde van het videofragment een hoge engagement (5) laat zien en gedurende 2 derde van het fragment een lagere engagement (3) laat zien. Dan is de uiteindelijke niveau voor engagement dus tussen een 3.5 en een 4. Handig is dus om tijdens de fragmenten te noteren hoe engaged het kind is en waarom je dat vindt. We gaan twee soorten engagement meten: taak engagement en robot engagement.

Het meten van taak engagement

Taak engagement kijkt naar hoe de kind engaged is met de taak. Dit kan op de tablet zijn, maar ook als de robot vraagt dat het kind iets moet doen (zoals nazeggen en nadoen). Als het kind doordat de robot praat richting de robot kijkt, is het kind nog steeds engaged met de taak. Ook in het geval dat het kind naar de robot kijkt als de robot een gebaar laat zien, leidt dit niet tot een lagere taak engagement. Immers, het nazeggen en de gebaren behoren tot de taak. Alleen in het geval dat het kind ergens anders op focust tijdens de taak of naar de robot kijkt zonder enige reden scoor je de taak engagement lager. Dit betekent ook dat je niet meet hoe engagement het kind met de robot is, dat is de focus van de andere engagement schaal.

Taak engagement gaat ook gepaard met fouten in het spel, over het algemeen leidt een lagere engagement tot meer fouten bij een kind. Maar, zoals jullie vast herkennen door het zelf afnemen van de experimenten, zag het systeem soms fouten die eigenlijk niet fout waren. In dit geval is het aan jou om deze fouten niet mee te laten tellen met jouw engagement score. Op de volgende pagina is de schaal voor taak engagement in een tabel met voorbeelden gezet.

Het meten van robot engagement

Robot engagement kijkt alleen naar hoe de kind engaged is met de robot. Dit is niet gerelateerd aan de taak. Het kind kan engaged met de robot zijn zonder dat het kind de taak uitvoert. Robot engagement wordt bepaald door de mate van hoe vaak het kind praat met de robot en kijkt richting de robot. Alleen het nazeggen van een target word is geen teken van robot engagement, immers de kinderen in de tablet conditie praten ook de tablet na. Als het kind bij het nazeggen van het target woord de robot ook nog aankijkt, dan telt het wel mee voor de robot engagement. Ook kinderen die de gebaren van de robot na doen laten een hoge engagement zien. Een kind dat alleen richting de tablet kijkt en de robot negeert (probeert te negeren) zal juist lager scoren.

Table 3: De schaal voor taak engagement

Niveau	Engagement	Voorbeelden
1	Uitgesproken laag	*Het kind vertoont nagenoeg geen activiteit:* - Geen concentratie: staren, wegdromen; - Een afwezige, passieve houding; - Geen gerichte activiteit, doelloze handelingen, niets teweegbrengen; - Alleen aan het tikken op het scherm om door te gaan - Alleen bezig met de experiment leider en niet met de taak; - Geen tekenen van exploratie en interesse; - Niets in zich opnemen, geen mentale activiteit.
2	Laag	- Het kind vertoont enige activiteit, maar deze wordt geregeld onderbroken: - Beperkte concentratie: wegkijken, prullen (friemelen), dromen; - Makkelijk afgeleid; - Taken worden in beperke mate uitgevoerd.
3	Matig	*Er is de hele tijd activiteit, maar niet echt geconcentreerd.* - Het kind is routinematig, vluchtig bezig; - Is beperkt gemotiveerd, voelt zich niet uitgedaagd, toont geen echte inzet; - Doet geen diepgaande ervaring op; - Is niet opgeslorpt door wat het doet; - Gebruikt zijn capaciteiten maar met mate; - De activiteit raakt de verbeelding en het denkvermogen van het kind niet. - De meeste taken worden uitgevoerd.
4	Hoog	*Er zijn doorgaans signalen van engagement:* - Het kind gaat globaal op in zijn spel; - Er is doorgaans concentratie, maar soms verslapt de aandacht - Het kind voelt zich uitgedaagd, er is een zekere gedrevenheid; - Gebruikt zijn capaciteiten; - Spreekt de verbeelding en het denkvermogen aan.
5	Uitgesproken hoog	*Het kind is gedurende de hele tijd ononderbroken bezig en gaat sterk op in zijn activiteit:* - Is ononderbroken geconcentreerd, opgeslorpt door de activiteit, vergeet de tijd; - Is heel gemotiveerd, voelt zich sterk aangesproken; - Is niet af te leiden; - Kijkt aandachtig naar de taak, heeft aandacht voor details; - Spreekt voortdurend al zijn capaciteiten en mogelijkheden aan; - Er is een sterke mentale activiteit: - De verbeelding en het denkvermogen draaien op volle toeren; - Doet diepgaande nieuwe ervaringen op; - Geniet van zo gedreven bezig te zijn.

Table 4: De schaal voor robot engagement

Niveau	Engagement	Voorbeelden
1	Uitgesproken laag	*Het kind vertoont nagenoeg geen interactie met de robot:* - Negeert de robot volledig; - Heeft een gesloten (lichaams)houding richting de robot; - Een afwezige, passieve houding; - Geen gerichte activiteit, doelloze handelingen, niets teweegbrengen; - Geen tekenen van interesse in de robot
2	Laag	*Het kind vertoont enige robot interactie, maar deze wordt geregeld onderbroken:* - Kijkt beperkt richting de robot; - Makkelijk afgeleid van de robot;
3	Matig	*Er is de hele tijd robot activiteit, maar niet echt geconcentreerd.* - Het kind is routinematig, vluchtig bezig; - Is beperkt gemotiveerd, voelt zich niet uitgedaagd, toont geen echte inzet; - Heeft een open (lichaams)houding richting de robot; - Is niet opgeslorpt door wat de robot doet; - Doelloos aanraken van de robot
4	Hoog	*Er zijn doorgaans signalen van robot engagement:* - Het kind gaat globaal op in zijn spel met de robot; - Er is doorgaans sprake van joint attention; - Er is doorgaans concentratie, maar soms verslapt de aandacht;
5	Uitgesproken hoog	*Het kind gaat sterk op in zijn activiteit met de robot:* - Is ononderbroken met de robot bezig - Is niet af te leiden van de robot; - Kijkt aandachtig naar robot, - Praat tegen de robot; - Gebaren na doen (in de iconische gebaren conditie); - Er is sprake van joint attention; - Geniet van met de robot bezig te zijn

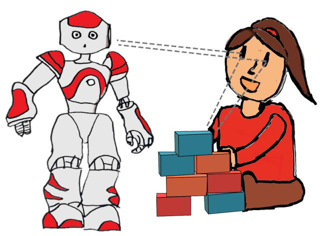
&

Summary

The impact of COVID-19 showed that our traditional classrooms are now heavily relying on digital tools. In the past few years (2020-2021), teachers had to teach children online and parents had to support their children in their school activities. Digital tools that can support during teaching, such as social robots, would have been extremely helpful for teachers. Robots have the advantage over tablets that they can use their body to act out behaviors similar to those of teachers. For instance, by accompanying speech with physical gestures that can help children remain focused and increase their learning gain. Moreover, this physical modality allows children to interact with robots more socially, which is especially important in second-language (L2) learning.

My PhD trajectory was part of the Horizon 2020 L2TOR project[1], in which six different universities and two companies worked together and investigated whether a humanoid robot can teach preschool children words from a second language. One of the key questions in this project was how we can develop robot behaviors that keep children engaged over time. Therefore, I conducted multiple studies to explore the effect of the robot on children's engagement and their perception of the robot.

The role of eye gaze in engagement



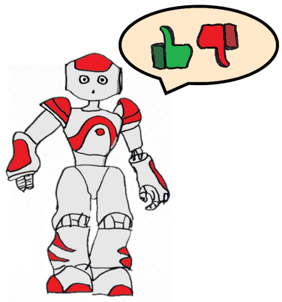
We first explored how we can measure children's engagement. We distinguished two types of engagement: engagement with the **task** and engagement with the **robot** in order to understand whether children were mostly captivated by the learning task, or the robot as interaction partner. To measure children's engagement, children's gaze direction is especially important, because it can show the direction of the children's attention. Therefore, we investigated the role of eye gaze in children's engagement with the task and the robot.

[1] The L2TOR project played a large role within the human-robot interaction (HRI) field toward open science. All of the L2tor publications, the project deliverables, source code and data have been made publicly available via the website `www.l2tor.eu` and via `www.github.nl/l2tor` and most studies were pre-registered.

In **Chapter two**, we found that the direction in which children looked during the experiment played a large role in children's engagement. Children's task engagement and robot engagement could largely be measured by the duration that children looked in a certain direction. For children's task engagement all gaze directions were important, and for children's robot engagement the most important gaze direction was toward the robot. However, both engagement types could not completely be measured by children's eye-gaze directions, which demonstrates that despite eye gaze having a large role in engagement, it does not predict all aspects of engagement. Therefore, we developed an engagement coding scheme that was used in the other studies.

Feedback

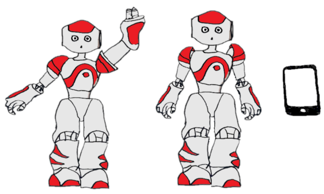
Using this coding scheme, we investigated what kind of robot behavior can influence these two engagement types. We found that the type of robotic feedback has an effect on task engagement and robot engagement but not on learning gain. In **Chapter two**, we compared the effect of two robotic feedback types on 3-year-old children's engagement and learning gains.



These feedback types were based on how adults provide feedback and how peers provide feedback. Moreover, we compared these two feedback strategies with a condition without any feedback. Contrary to our expectations, we observed no differences in this study regarding children receiving feedback or no feedback. We suspected that feedback may have more impact when provided over a longer time period.

Consequently in **Chapter three**, we investigated 5-year-old children's engagement, who have a longer attention span than 3-year-old children. We also designed three lessons instead of one. In addition, we wanted to make sure that we designed the feedback to be as effective as possible and, thus, we interviewed student teachers how they would provide feedback using the robot. We used their answers in a teacher-preferred feedback strategy and compared it to the, as discussed with them, worst feedback strategy. This time, children learned and the robot sessions were successful in teaching children new vocabulary. However, children learned as many words with the robot's feedback as without feedback. We did find a difference in children's engagement. Children were more engaged with the robot and task when interacting with a robot using teacher-preferred feedback than in the other conditions. Moreover, some children mentioned the feedback in an after experiment interview, where they said that they liked that the robot helped them finding the correct answer.

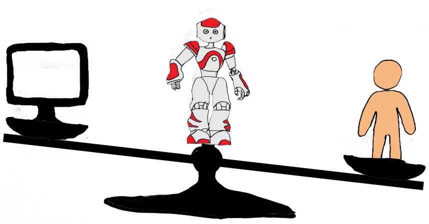
Gestures



As explained before, the type of feedback can have an influence on both children's task engagement and robot engagement. However, we did not yet find a difference between children's task engagement and robot engagement. To explore the added value of a robot compared to the task, **Chapter four** describes a study in which we investigated a robot using iconic gestures (gestures that depict the meaning of a second-language concept), a robot using no iconic gestures, compared to a tablet-only condition.

To explore this effect on the long-term, we conducted a seven sessions study with over 200 children. The task was the same in all conditions, in order to have consistent results. This resulted in no differences in children's task engagement. However, when looking at children's robot engagement, we found that children were more robot-engaged when the robot used iconic gestures than with a robot without iconic gestures. It seems that the robot gestures had a positive influence on children's robot engagement. A robot in the iconic gesture condition that moves its arms and hands will draw more attention to itself, resulting in a *higher* robot engagement. These findings imply that the robot's gestures must be well designed in an interaction.

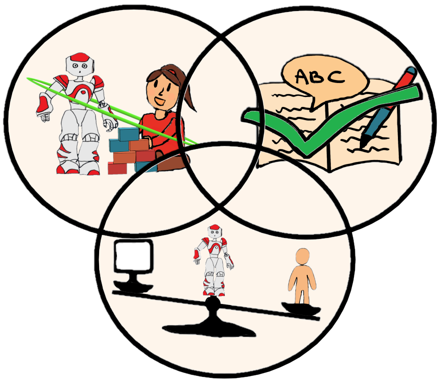
Perception



Finally, by interacting with robots over a long period of time, children will undoubtedly develop a relationship with the robot. This relationship can change over time, due to the behavior of the robot.

Therefore, in **Chapter five**, we investigated the children's perception prior to the seven sessions study, and after this study. We found that children perceived the robot more as a human before the experiment than after the experiment and that specifically boys perceived the robot more as a computer after the experiment. We suspect that the reason for this is because the children had very high expectations before the experiment, ones that the robot was not able to fulfill and therefore, children's perception of the robot changed after the experiment.

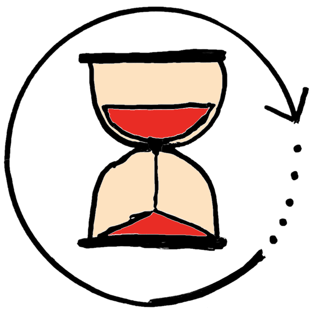
Relation engagement, perception and word learning



Over the different chapters, we investigated the relation between children's engagement and their word knowledge after the experiment. Word knowledge was related to a higher engagement, but this held true for both types of engagement and only in our study during seven sessions. This study showed that children's word knowledge was higher when children were engaged with the task, or engaged with the robot, or both. This also holds the other way around, children who remembered fewer words were less engaged with the task and the robot.

Moreover, in **Chapter five**, we also explored the relation between children perception of the robot and their learning. We found that children's perception before the experiment was related to their word knowledge after the lessons. But also that children who viewed the robot more as a human over time, remembered more words over time and vice versa, children who perceived the robot more as a computer after the experiment, remembered less words over time.

Long-term effects



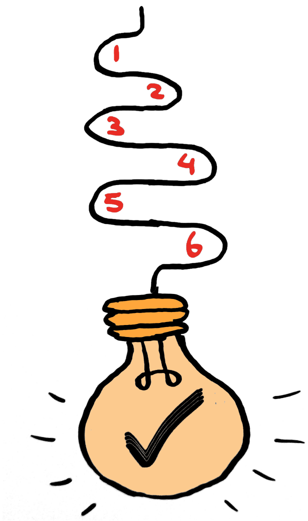
Engagement is not only dependent on the robot's behavior, it also is dependent on time. For this purpose, we measured children's engagement over multiple sessions (in **Chapters three and four**). Children's engagement dropped over time during both experiments, but there were large individual differences. Moreover, the robot's behavior and the content of the lessons have an influence on the decrease in children's engagement.

Chapter three showed that the robot's feedback strategy in one session had an influence on the children's engagement in the following session. Furthermore, we noticed that during our seven sessions study in **Chapter four**, after the third lesson, children's engagement

&

seemed to decrease less (although it still varied for each child and lesson). Moreover, we found that the final session increased children's engagement again, possibly because this session had a different setup than the rest. It is therefore important to take previous interactions into account when designing long-term child-robot interactions and to expect that children's engagement will first decrease and will later stabilize over time.

Concluding remarks



This dissertation expands our knowledge for using robots to support children in second-language learning. A robot tutor would have been extremely valuable during the COVID-19 pandemic, but even after the pandemic this trend may continue. However, before being able to introduce the robot in education and making the future classroom a reality, it is important to take further steps. On the one hand, it is good to already deploy robots for tasks they are currently suited for, such as single lessons tailored to maximize the contribution of robots. On the other hand, for the future, more studies need to be carried out over longer periods to explore children's engagement. With these extra experiments in mind, we are confident that robots will increasingly be able to support the teachers of the future.

Samenvatting

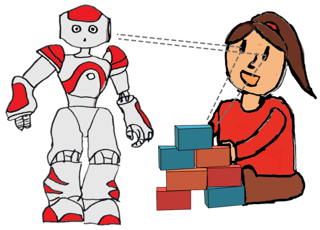
Covid-19 heeft laten zien dat onze traditionele manier van lesgeven steeds meer afhankelijk is van digitale hulpmiddelen. In de afgelopen jaren (2020-2021) hebben leerkrachten kinderen online les moeten geven en hebben ouders hun kinderen moeten begeleiden bij hun lesactiviteiten. Digitale instrumenten die het onderwijs kunnen ondersteunen zoals sociale robots, zouden uiterst nuttig zijn geweest voor leerkrachten. Robots die, in tegenstelling tot tablets, hun lichaam kunnen gebruiken om zich vergelijkbaar te gedragen als leerkrachten. Bijvoorbeeld door te gebaren tijdens het praten, waardoor kinderen zich beter kunnen concentreren wat een voordeel oplevert voor hun leerprestaties. Bovendien stellen robots, meer dan tablets, kinderen in staat tot een sociale interactie, wat vooral belangrijk is bij het leren van een tweede taal (L2).

Hierover ging mijn promotietraject wat onderdeel was van het Horizon 2020 L2TOR project[2], waarin zes verschillende universiteiten en twee bedrijven samenwerkten en onderzochten of een robot aan kleuters woorden uit een tweede taal kon leren. Een van de belangrijkste vragen in dit project was hoe we gedrag van de robot konden ontwikkelen dat kinderen betrokken (engaged) houdt. Betrokkenheid van kinderen is belangrijk zodat zij tijdens langere tijdsperiodes met de robot aan de slag willen. Om deze vraag te beantwoorden, heb ik meerdere studies uitgevoerd om het effect van de robot op de betrokkenheid van kinderen met de robot te onderzoeken, alsmede onderzoek te doen naar de perceptie die de kinderen van de robot hadden.

[2] Het L2TOR project leverde een grote bijdrage binnen het mens-robot interactie veld in de beweging richting publieke wetenschap. Alle L2TOR publicaties, de project deliverables, broncode en data zijn openbaar gemaakt via de website `www.l2tor.eu` en via `www.github.nl/l2tor` en de meeste studies werden vooraf geregistreerd.

&

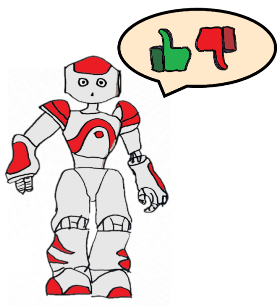
De rol van kijkgedrag in betrokkenheid



We hebben eerst onderzocht hoe we de betrokkenheid van kinderen kunnen meten. We maakten onderscheid tussen twee vormen van betrokkenheid: enerzijds met de **leertaak** zelf en anderzijds in de omgang met de **robot**; dit was om te bepalen of kinderen vooral geboeid waren vanwege de leertaak, of vanwege de robot. Om de betrokkenheid van kinderen te bepalen kan hun kijkrichting belangrijk zijn, omdat deze aangeeft in welke richting de aandacht van de kinderen gaat. Daarom onderzochten we de rol van kijkgedrag in betrokkenheid bij de taak en de robot.

In hoofdstuk 2 vonden we dat de richting waarin kinderen keken tijdens het experiment een grote rol speelde in de mate waarin ze betrokken waren. Betrokkenheid met de taak en met de robot werden grotendeels gemeten door de duur dat kinderen in een bepaalde richting keken. Voor de betrokkenheid van kinderen bij de leertaak waren alle kijkrichtingen belangrijk, en voor de betrokkenheid met de robot was het kijken in de richting van de robot het meest belangrijk. Beide vormen van betrokkenheid konden echter niet volledig worden gemeten door het kijkgedrag van de kinderen, wat aantoont dat hoewel de kijkrichting een grote rol speelt in betrokkenheid, het niet alle aspecten van betrokkenheid voorspelt. Daarom ontwikkelden we een coderingsschema om betrokkenheid te meten om te gebruiken in de overige studies.

Feedback

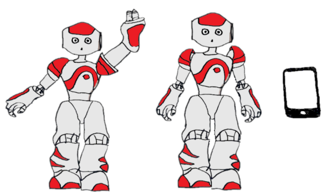


Met behulp van dit coderingsschema, hebben we onderzocht welk gedrag van de robot deze twee soorten van betrokkenheid kan beïnvloeden. We vonden dat het type feedback dat de robot gebruikte een effect had op zowel de betrokkenheid van het kind bij de taak als op de betrokkenheid in de omgang met de robot, maar niet op het leerresultaat.

In **hoofdstuk 2** vergeleken we het effect van twee vormen van feedback op betrokkenheid en leerresultaat van driejarige kinderen. De robot gaf ofwel feedback zoals volwassenen feedback geven, ofwel hoe leeftijdsgenoten feedback geven. Bovendien vergeleken we deze twee feedbackmethoden met een robot die geen feedback gaf. In tegenstelling tot onze verwachtingen, zagen we geen verschillen tussen kinderen die een van de twee verschillende feedback of geen feedback kregen. We vermoedden dat feedback meer effect zou kunnen hebben wanneer deze gedurende een langere periode wordt gegeven.

Daarom hebben we in **hoofdstuk 3**, ook de betrokkenheid van vijfjarige kinderen onderzocht, die een langere aandachtsspanne hebben dan driejarige kinderen, en bovendien ontwierpen we drie lessen in plaats van één, zodat de feedback over een langere periode zou worden gegeven. Daarbij wilden we er zeker van zijn dat de feedback die de robot zou geven zo effectief mogelijk door ons zou worden ontworpen. Daarvoor hebben we studenten van de lerarenopleiding gevraagd hoe zij feedback zouden geven als zij de robot zouden gebruiken. We gebruikten hun antwoorden in de 'door de leerkracht uitgekozen' feedbackmethode en vergeleken die met de, zoals met hen besproken, 'slechtste' feedbackmethode. Deze keer zagen we dat de robotsessies succesvol waren in het aanleren van nieuwe woorden aan kinderen. Kinderen leerden echter evenveel woorden met de robot's feedback als zonder feedback. We vonden wel een verschil in de betrokkenheid van kinderen. Kinderen waren meer betrokken met de robot en bij de taak wanneer ze interactie hadden met een robot die de 'door de leraar uitgekozen' feedback gebruikte dan in de andere condities. Bovendien gaven sommige kinderen in een interview na afloop van het experiment aan dat ze het fijner vonden dat de robot hen hielp bij het vinden van het juiste antwoord.

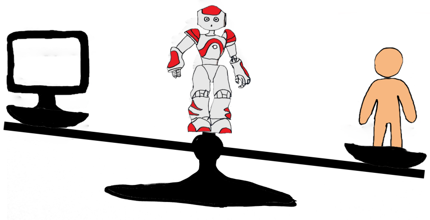
Gebaren



Zoals besproken kan het type feedback invloed hebben op de betrokkenheid van kinderen bij zowel de taak als met de robot. We zagen echter nog geen toegevoegde waarde van de robot ten opzichte van de taak. Daarom onderzochten we in **hoofdstuk 4** de verschillen tussen het uitvoeren van een taak met een robot met iconische gebaren (gebaren die een link maken tussen het tweedetaalwoord en de betekenis), het uitvoeren van een taak met een robot zonder iconische gebaren, en dezelfde taak maar dan alleen met een tablet uitgevoerd.

Om dit effect op lange termijn te onderzoeken, voerden wij een zeven lessen onderzoek uit met meer dan 200 kinderen. Het lessenprogramma was onder alle drie condities hetzelfde zodat de deze goed vergelijkbaar waren. Hierdoor waren gemiddeld alle kinderen even betrokken waren bij de taak. Toen we echter naar de betrokkenheid van de kinderen met de robot keken, vonden we dat kinderen meer betrokken waren met een robot die iconische gebaren gebruikte dan met een robot zonder iconische gebaren. Het lijkt erop dat deze gebaren een positieve invloed hadden op hoe betrokken kinderen met de robot waren. Een robot met iconische gebaren beweegt zijn armen en handen, trekt meer aandacht naar zich toe, wat resulteert in een hogere betrokkenheid. Deze bevindingen suggereren dat de gebaren van de robot goed ontworpen moeten zijn in een interactie.

Perceptie

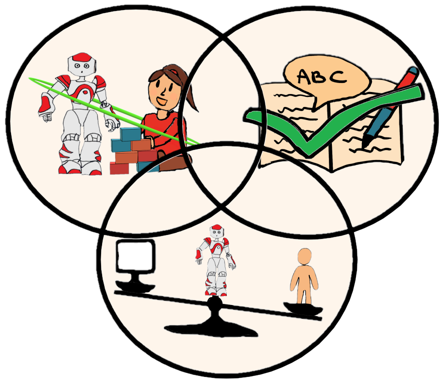


Ten slotte zullen kinderen, door gedurende lange tijd met robots om te gaan, onvermijdelijk een relatie met de robot ontwikkelen. Deze relatie kan in de loop van de tijd veranderen, door het gedrag van de robot.

Daarom onderzochten we in ons laatste onderzoek, in **hoofdstuk 5**, de perceptie van de kinderen voorafgaand aan het zeven lessen onderzoek, en na afloop van de lessen. We ontdekten dat kinderen de robot vóór het experiment meer als een mens zagen dan na het experiment. Vooral jongens gingen de robot meer als een computer zien. De verandering van perceptie kan komen doordat de kinderen vóór het experiment zeer hoge verwachtingen hadden, die de robot niet kon waarmaken, en dat daarom kinderen hun perceptie van de robot na het experiment veranderden.

Relatie betrokkenheid, perceptie en leerprestaties

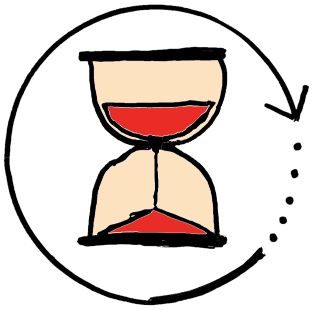
In de verschillende studies onderzochten we de relatie tussen de betrokkenheid van de kinderen en hun woordkennis na het experiment. Er bleek een verband te bestaan tussen woordkennis en een hogere betrokkenheid, maar we vonden dit alleen in onze zeven lessen studie. Dit verband gold voor zowel betrokkenheid met de robot als bij de taak: onze studie toonde aan dat de kinderen meer woorden onthielden wanneer kinderen betrokken waren bij de taak, of betrokken met de robot, of beide.



Dit gold ook de andere kant op, kinderen die minder woorden onthielden waren minder betrokken met de taak en de robot. Daarnaast onderzochten we in **hoofdstuk 5** het verband tussen de perceptie van de robot en hun leerprestaties. We toonden aan dat de perceptie die de kinderen vóór het experiment van de robot hadden, gerelateerd was aan hun woordkennis na de lessen. We ontdekten ook dat kinderen die de robot na verloop van tijd meer als een mens zagen, meer woorden onthielden. Vice versa, de kinderen die de robot na het experiment meer als een computer zagen, onthielden minder woorden na verloop van tijd.

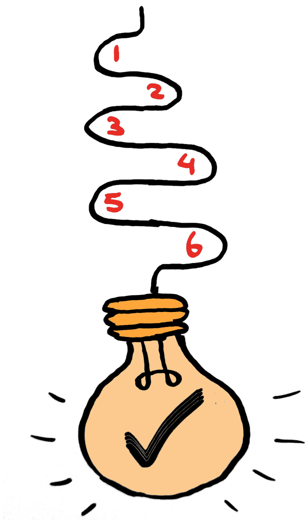
Lange termijn effecten

Beide types van betrokkenheid zijn niet alleen afhankelijk van het gedrag van de robot, maar ook van tijd. Daarom hebben we de betrokkenheid van kinderen gemeten gedurende de experimenten in **hoofdstukken 3 en 4** met drie en zeven lessen. De betrokkenheid van kinderen daalde in de loop van de tijd tijdens beide experimenten, maar er waren grote individuele verschillen. Bovendien hebben het gedrag van de robot en de inhoud van de lessen invloed op de daling van de betrokkenheid van kinderen.



Uit **hoofdstuk 3** bleek dat de feedbackmethode van de robot in een les invloed had op de betrokkenheid van kinderen tijdens de volgende les. Verder bleek dat bij het onderzoek van **hoofdstuk 4**, de betrokkenheid van kinderen na de derde les minder sterk leek af te nemen (hoewel het nog steeds per kind en les verschilde). Bovendien ontdekten we dat de betrokkenheid van kinderen tijdens de laatste les weer omhoogging, mogelijk omdat deze les woorden herhaalden en daarmee een andere opbouw had dan de overige lessen. Het is daarom belangrijk om rekening te houden met eerdere interacties bij het ontwerpen van lange termijn kind-robot interacties en ervan uit te gaan dat hun betrokkenheid over tijd zal afnemen, maar uiteindelijk op een bepaald niveau zal blijven hangen.

Tot slot



Dit proefschrift draagt bij aan kennis over het gebruik van robots om kinderen te ondersteunen bij het leren van een tweede taal. Een robot tutor zou zeer waardevol zijn geweest tijdens de Covid-19 pandemie, maar ook na de pandemie kan deze tendens zich voortzetten. Voordat we de robot in het onderwijs kunnen introduceren en we het toekomstige klaslokaal werkelijkheid kunnen maken, is het belangrijk om verdere stappen te zetten. Enerzijds is het goed om robots nu al in te zetten voor taken waar ze op dit moment geschikt voor zijn, zoals losse lessen optimaal ontworpen voor de robot. Anderzijds moeten er voor de toekomst meer studies worden uitgevoerd over langere periodes om meer inzicht te krijgen in de betrokkenheid van kinderen. Met deze extra experimenten in gedachten hebben wij er alle vertrouwen in dat robots in toenemende mate in staat zullen zijn de leraren van de toekomst te ondersteunen.

Acknowledgments

De afgelopen jaren heb ik echt genoten van het promotietraject en onderstaande mensen zijn daar mede verantwoordelijk voor. Ik wil hen daarom van het diepste van mijn hart bedanken.

Paul en Emiel, kan ik eindelijk mijn 2-cent input geven. Jullie input was sowieso euros waard, ik had dit echt niet zonder jullie kunnen doen. Ik waardeer jullie betrokkenheid enorm en jullie kritische blikken en enthousiasme hebben mijn phd-carrière laten groeien op persoonlijk en professioneel vlak. Maar als ik terugdenk denk ik vooral aan de bijzondere momenten die we hebben meegemaakt. Bijvoorbeeld het moment dat onze eerste robots binnenkwamen en we ze gingen uitpakken en meteen alle dansjes uitprobeerden, het buiten gesloten zijn van ons hotel in Wenen, en natuurlijk ons experiment op Lowlands. Heel erg bedankt dat jullie mij toendertijd aangenomen hebben op het L2TOR project.

I also want to thank **Emilia, Tibor, Koen and Pieter S**, who were willing to accept the task of reviewing my dissertation, providing suggestions, and being present at my PhD defence. Emilia, thank you for starting my 'research spark' all these years ago when I did my internship at the TU/E.

Rianne B en Rianne C, ik ben vereerd dat jullie aan mijn zijde staan tijdens mijn verdediging als mijn twee paranymfen. En wat handig dat ik gewoon Rianne kan zeggen en ik twee personen ter beschikking heb. Rianne Conijn, er was niemand anders met wie ik ons wortelonderzoek had willen doen. Ik kon altijd heerlijk met jou praten over alles waar we tegenaanliepen binnen de universiteit, dit met regelmaat met een rondje bossen tijdens de lunch. Rianne B, je was mijn maatje om mee te sparren over onderzoek (nu nog steeds) ondanks dat je een hele andere achtergrond hebt, of wellicht juist daarom. Ons verblijf in Lausanne, waar we naast een kamer ook schoenen deelden zal altijd mij bij blijven.

Jan en Bram, onze kantoor was altijd gezellig, en we stonden erom bekend dat wij altijd snoep of chocolade hadden om uit te delen. De eerste keer sushi eten was met jullie in Wenen, en daarna stond Jan altijd met zijn auto klaar stond om ergens heen te rijden, danwel naar de ene sushi bar in Tilburg, danwel naar de andere sushi bar in Tilburg, of een Japans restaurant in Parijs. Ik wacht nog op het moment dat we naar Japan gaan, we hebben al flink geoefend. Moving on to the rest of the **L2TOR team**, I also want to thank them. I am so happy that we were not only hired on the same project, but we were actually a team working together and

making the first steps in this new field, including great hackathons in Paris at the Softbank head quarters and drinking beers in Eindhoven. Besonderen Dank an **Thorsten**, with who my German practice was short-lived. Daarnaast hebben er veel studenten meegewerkt binnen onze projecten. Bedankt **Annabella, Chani, Chiara, Hugo, Esmee, Laura, Laurette, Madee, Marije, Michelle, Peta, Reinjet, Robin S, Sabine, Sam, Sirkka** voor jullie hulp met het verzamelen van de data. **Pieter W**, bedankt dat ik jou mocht begeleiden tijdens je afstuderen en heel veel succes met jouw promotie!

Ook dank aan velen van de universiteit (I hope I didn't forget anyone). **Alexandra, Anna, Angelique, Chris, George, Lieke, Lisa, Mariana, Marlies, Menno Z, Paris, Thiago and Tycho**, you showed that enjoyment and productivity go together and that you need breaks and especially Friday evening drinks. **Pieter S en Marie**, ik wil jullie bedanken voor de gelegenheid om aan mijzelf te blijven werken tijdens mijn promotietraject zodat ik alvast onderwijservaring op kon doen. Marie, bedankt ook het aanmoedigen om mijn eigen tekeningen te gebruiken in mijn proefschrift. **Gonzalo, Maryam, Merel, Lisanne, Peter**, I want to thank you that you have welcomed me in your team of Assistant Profs. **Dimitar, Mirella**, let's make the robot do sign language at some point. **Eva en Karin**, heel erg bedankt voor al jullie hulp de afgelopen jaren. Natuurlijk kan het ook niet achterblijven om **Robin, Luka, Charlie, en Taylor** te bedanken, zonder hen was dit niet mogelijk geweest. Ik ben heel blij dat we er heelhuids doorheen gekomen zijn (op een paar gekneusde vingers na)!

Ik heb ook velen buiten het werk om gehad om afleiding te geven. **Irene**, we hebben wat meegemaakt in ons kleine appartementje in Nijmegen, maar ik had nooit voorspeld dat ik zou verhuizen naar de plek waar jij ongeveer ben opgegroeid. Ik heb altijd heerlijk kunnen praten met jou over de kids, en hopelijk kan ik nog een keertje langskomen met de robot! **Iris, Kimberley, Sigrid, Ashley, Monique en Daria**, ik wil jullie bedanken voor het plezier buiten mijn werk om als ik weer energie teveel had gekregen van de kinderen en het eraf moest sporten, of als ik juist helemaal in dubio stond hoe ik verder moest gaan. **Maarten en Leonoor**, jullie bedankt voor alle afleiding met spelletjes spelen en het borrelen op maandagavonden om de eerste werkdag van de week af te sluiten.

Papa en mama, bedankt dat jullie mij altijd hebben aangemoedigd in wat ik wilde doen. Als ik nog even een halfjaartje naar het buitenland wilde ondanks dat dat vertraging zou opleveren, als ik wilde verhuizen naar de andere kant van het land (Noord-Brabant), niets is te ver voor jullie en jullie hebben mij altijd geholpen wanneer ik het vroeg (maar ook als ik het niet vroeg). Jullie begrepen mijn academische dilemma's precies en ik kwam altijd verder als ik met jullie sprak. Ik ben vandaag extra blij dat mijn promotie nog net voor papa's pensioen komt.

Wouter en Menno, bedankt dat jullie mij hebben voorbereid op het peer review proces door jullie plagerijen vroeger maar ook voor alle hoeraatjes tegenwoordig als ik door het proces heengegaan ben. **Arjen**, natuurlijk wil ik ook jou niet vergeten. Jij hebt van heel dichtbij meegemaakt hoe zwaar ik het af en toe had, en dat het lastig was om alles hoog te houden. Maar jij zorgde er altijd voor dat het duidelijk was dat het af en toe niet erg is om een balletje te laten vallen en tijd voor mijzelf (en ons) te nemen. Heel erg bedankt daarvoor. Ik wil dan ook je ouders even noemen, want niemand was blijer voor mij dan jullie, **Ad en Elly**, toen ik mijn werk indiende.

Als laatste wil ik graag alle scholen, kinderen, ouders, juffen en meesters bedanken die meegedaan hebben aan mijn onderzoeken. In de afgelopen jaren heb ik meer dan 300 kinderen gezien en getest, de meesten meerdere weken en dat verdient zeker een plek in mijn dankwoord.

&

List of publications

Journal papers (peer reviewed)

- **de Haas, M.**, Vogt, P., & Krahmer, E. (2021). When preschoolers interact with an educational robot, does robot feedback influence engagement? *Multimodal Technologies and Interaction, 5(12)*. doi: 10.3390/mti5120077
- van den Berghe*, R., **de Haas*, M.**, Oudgenoeg-Paz, O., Krahmer, E., Verhagen, J., Vogt, P.,... Leseman, P. (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. doi: https://doi.org/10.1111/jcal.12497
- Leeuwestein, H., Barking, M., Sodacı, H., Oudgenoeg-Paz, O., Verhagen, J., Vogt, P., ... **de Haas, M.**,... Leseman, P. (2021). Teaching Turkish-Dutch kindergartners Dutch vocabulary with a social robot: Does the robot's use of Turkish translations benefit children's Dutch vocabulary learning? *Journal of Computer Assisted Learning, 37(3)*, 603–620.
- van den Berghe, R., Oudgenoeg-Paz, O., Verhagen, J., Brouwer, S., **de Haas, M.**, de Wit, J., ... Leseman, P. (2021). Individual differences in children's (language) learning skills moderate effects of robot-assisted second-language learning. *Frontiers in Robotics and AI* , 8, 259. doi: 10.3389/frobt.2021.676248
- **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. Frontiers in Robotics and AI , 7, 101. doi: 10.3389/frobt.2020.00101
- **de Haas, M.**, Aroyo, A. M., Haselager, P., Smeekens, I., & Barakova, E. (2018). Comparing robots with different levels of autonomy in educational setting. In *Practical issues of intelligent innovations* (pp. 293–311). Springer.
- Barakova, E. I., **de Haas, M.**, Kuijpers, W., Irigoyen, N., & Betancourt, A. (2018). Socially grounded game strategy enhances bonding and perceived smartness of a humanoid robot. *Connection Science, 30(1)*, 81–98.
- Belpaeme, T., Vogt, P., van den Berghe, R., Bergmann, K., Göksun, T., **de Haas, M.**, ... others (2018). Guidelines for designing social robots as second language tutors. *International Journal of Social Robotics, 10(3)*, 325-341.

- **de Haas, M.**, Smeekens, I., Njeri, E., Haselager, P., Buitelaar, J., Lourens, T., ... Barakova, E. (2017). Personalizing educational game play with a robot partner. In *Robotics in education* (pp. 259–270). Springer.
- Vogt, P., **de Haas, M.**, de Jong, C., Baxter, P., & Krahmer, E. (2017). Child-robot interactions for second-language tutoring to preschool children. *Frontiers in Human Neuroscience, 11*. doi: 10.3389/fnhum.2017.00073

Journal papers (under review)

- **de Haas, M.**, Vogt, P., van den Berghe, R., Leseman, P., Oudgenoeg-Paz, O., Willemsen, B., ... Krahmer, E. (under review). Engagement in longitudinal child-robot language learning interactions: Disentangling robot and task engagement.

Other chapters

- **de Haas, M.**, Vogt, P., & Krahmer, E. (2016a). Taal leren met behulp van een sociale robot. *DIXIT*.

Conference contributions in proceedings (peer reviewed)

- **de Haas, M.**, & Conijn, R. (2020). Carrot or stick: The effect of reward and punishment in robot assisted language learning. In *Companion of the 2020 acm/ieee international conference on human-robot interaction* (p. 177–179). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3371382.3378349
- Zhang, P., & **de Haas, M.** (2020). Effects of pitch gestures on learning chinese orthography with a social robot. In *Companion of the 2020 acm/ieee international conference on human-robot interaction* (p. 180–182). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3371382.3378350
- de Wit, J., Willemsen, B., **de Haas, M.**, Krahmer, E., Vogt, P., Merckens, M., ... Wolfert, P. (2019). Playing charades with a robot: Collecting a large dataset of human gestures through hri. In *Companion of the 2019 acm/ieee international conference on human-robot interaction* (pp. 634–635).
- Vogt, P., van den Berghe, R., **de Haas, M.**, Hoffman, L., Kanero, J., Mamus, E., ... others (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). doi: 10.1109/HRI.2019.8673077

- de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., **de Haas, M.**, Kopp, S., ... Vogt, P. (2018). The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (pp. 50–58). Association for Computing Machinery.
- Willemsen, B., de Wit, J., Krahmer, E., **de Haas, M.**, & Vogt, P. (2018). Context-sensitive natural language generation for robot-assisted second language tutoring. In *Proceedings of the workshop on nlg for human–robot interaction* (pp. 1–7).
- Wolfert, P., **de Haas, M.**, Vogt, P., & Haselager, P. (2018). Measuring engagement online in cri. In *Acm sigchi idc2018 workshop on children's robotics and child-robot interaction.*
- **de Haas, M.**, Baxter, P., de Jong, C., Krahmer, E., & Vogt, P. (2017). Exploring different types of feedback in preschooler and robot interaction. In *Proceedings of the companion of the 2017 acm/ieee international conference on human-robot interaction* (p. 127–128). New York, NY, USA: Association for Computing Machinery.
- Baxter, P., de Jong, C., Aarts, R., **de Haas, M.**, & Vogt, P. (2017). The effect of age on engagement in preschoolers' child-robot interactions. In *Proceedings of the companion of the 2017 acm/ieee international conference on human-robot interaction* (p. 81–82). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3029798.3038391
- Kanero, J., **de Haas, M.**, Mamus, E., Oranç, C., van den Berghe, R., Verhagen, J., ... others (2017). *Observing human tutoring to develop robot-based language lessons.* (Symposium on Multimodal Communication 2017, Bielefeld, Germany)
- Johal, W., Vogt, P., Kennedy, J., **de Haas, M.**, Paiva, A., Castellano, G., ... Dillenbourg, P. (2017). Workshop on robots for learning: R4l. In *Proceedings of the companion of the 2017 acm/ieee international conference on human-robot interaction* (pp. 423–424).
- Vogt, P., de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., **de Haas, M.**, ... Krahmer, E. (2017). Adaptation and gestures in second language tutoring using social robots. (Workshop on Early Literacy and (Digital) Media)
- **de Haas, M.**, Vogt, P., & Krahmer, E. (2016b). Enhancing child-robot tutoring interactions with appropriate feedback. In *Proceedings of the Long-Term Child-Robot Interaction Workshop at RO-MAN 2016.*
- **de Haas, M.**, Aroyo, A. M., Barakova, E., Haselager, P., & Smeekens, I. (2016). The effect of a semi-autonomous robot on children. In *Proceedings of the 2016 ieee 8th international conference on intelligent systems (is)* (pp. 376–381).

SIKS dissertation series

2016

1. Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
2. Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
3. Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
4. Laurens Rietveld (VU), Publishing and Consuming Linked Data
5. Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
6. Michel Wilson (TUD), Robust scheduling in an uncertain environment
7. Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
8. Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
9. Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10. George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11. Anne Schuth (UVA), Search Engines that Learn from Their Users
12. Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13. Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14. Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15. Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16. Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17. Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18. Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19. Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20. Daan Odijk (UVA), Context & Semantics in News & Web Search
21. Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22. Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23. Fei Cai (UVA), Query Auto Completion in Information Retrieval
24. Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25. Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26. Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27. Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28. Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29. Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30. Ruud Mattheij (UvT), The Eyes Have It
31. Mohammad Khelghati (UT), Deep web content monitoring
32. Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33. Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34. Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35. Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36. Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37. Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38. Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39. Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

&

40. Christian Detweiler (TUD), Accounting for Values in Design
41. Thomas King (TUD), Governing Governance: A formal framework for analysing institutional design and enactment governance
42. Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43. Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44. Thibault Sellam (UVA), Automatic Assistants for Database Exploration
45. Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46. Jorge Gallego Perez (UT), Robots to Make you Happy
47. Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48. Tanja Buttler (TUD), Collecting Lessons Learned
49. Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50. Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

1. Jan-Jaap Oerlemans (UL), Investigating Cybercrime
2. Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
3. Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
4. Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
5. Mahdieh Shadi (UVA), Collaboration Behavior
6. Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
7. Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
8. Rob Konijn (VU), Detecting interesting differences:data mining in health insurance data using outlier detection and subgroup discovery
9. Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10. Robby van Delden (UT), (Steering) Interactive Play Behavior
11. Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12. Sander Leemans (TUE), Robust Process Mining with Guarantees
13. Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14. Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15. Peter Berck (RUN), Memory-Based Text Correction
16. Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines
17. Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18. Ridho Reinanda (UVA), Entity Associations for Search
19. Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20. Mohammadbashir Sedighi (TUD), Fostering engagement in knowledge sharing: the role of perceived benefits, costs and visibility
21. Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22. Sara Magliacane (VU), Logics for causal inference under uncertainty
23. David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24. Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25. Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26. Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27. Michiel Joosse (UT), Investigating positioning and gaze behaviors of social robots: people's preferences, perceptions and behaviors
28. John Klein (VU), Architecture Practices for Complex Contexts
29. Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"
30. Wilma Latuny (UvT), The Power of Facial Expressions
31. Ben Ruijl (UL), Advances in computational methods for QFT calculations
32. Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33. Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34. Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35. Martine de Vos (VU), Interpreting natural science spreadsheets
36. Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37. Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38. Alex Kayal (TUD), Normative Social Applications
39. Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40. Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41. Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42. Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43. Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44. Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45. Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46. Jan Schneider (OU), Sensor-based Learning Support
47. Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48. Angel Suarez (OU), Collaborative inquiry-based learning

2018

1. Han van der Aa (VUA), Comparing and Aligning Process Representations
2. Felix Mannhardt (TUE), Multi-perspective Process Mining
3. Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
4. Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
5. Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
6. Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
7. Jieting Luo (UU), A formal account of opportunism in multi-agent systems
8. Rick Smetsers (RUN), Advances in Model Learning for Software Systems
9. Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10. Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology
11. Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks
12. Xixi Lu (TUE), Using behavioral context in process mining
13. Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14. Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters
15. Naser Davarzani (UM), Biomarker discovery in heart failure
16. Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17. Jianpeng Zhang (TUE), On Graph Sample Clustering
18. Henriette Nakad (UL), De Notaris en Private Rechtspraak
19. Minh Duc Pham (VUA), Emergent relational schemas for RDF
20. Manxia Liu (RUN), Time and Bayesian Networks
21. Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games
22. Eric Fernandes de Mello Araujo (VUA), Contagious: modeling the spread of behaviours, perceptions and emotions in social networks
23. Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24. Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis
25. Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
26. Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
27. Roelof Anne Jelle de Vries (UT), Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
28. Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29. Yu Gu (UVT), Emotion Recognition from Mandarin Speech
30. Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019

1. Rob van Eijk (UL), Comparing and Aligning Process Representations
2. Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
3. Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
4. Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
5. Sebastiaan van Zelst (TUE), Process Mining with Streaming Data
6. Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
7. Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
8. Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
9. Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems
10. Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11. Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12. Jacqueline Heinerman (VU), Better Together
13. Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14. Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15. Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16. Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17. Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18. Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19. Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20. Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21. Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22. Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture
23. Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24. Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25. Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description
26. Prince Singh (UT), An Integration Platform for Synchromodal Transport
27. Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses
28. Esther Kuinderman (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29. Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30. Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31. Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics
32. Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33. Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34. Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and
35. Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming
36. Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills
37. Jian Fang (TUD), Database Acceleration on FPGAs
38. Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020

1. Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
2. Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
3. Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
4. Maarten van Gompel (RUN), Context as Linguistic Bridges
5. Yulong Pei (TUE), On local and global structure mining
6. Preethu Rose Anish (UT), Stimulation architectural thinking during requirements elicitation - an approach and tool support

7. Wim van der Vegt (OUN), Towards a software architecture for reusable game components
8. Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
9. Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
10. Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining
11. Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
12. Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
13. Marco Virgolin (CWI), Design and application of gene-pool optimal mixing evolutionary algorithms for genetic programming
14. Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
15. Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
16. Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
17. Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
18. Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
19. Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
20. Albert Hankel (VU), Embedding Green ICT Maturity in Organisations
21. Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be
22. Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
23. Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
24. Lenin da Nobrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
25. Xin Du (TUE), The Uncertainty in Exceptional Model Mining
26. Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27. Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28. Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29. Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30. Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst
31. Gongjin Lan (VU), Learning better – From Baby to Better
32. Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising
33. Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation
34. Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35. Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021

1. Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
2. Rijk Mercuur (TUD), Simulating Human Routines:Integrating Social Practice Theory in Agent-Based Models
3. Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices
4. Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
5. Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
6. Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
7. Armel Lefebvre (UU), Research data management for open science
8. Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
9. Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10. Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11. Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12. Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13. Bob R. Schadenberg (UT), Robots for autistic children: understanding and facilitating predictability for engagement in learning
14. Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15. Onat Ege Adali (TU/e), Transformation of value propositions into resource re-configurations through the business services paradigm
16. Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

&

17. Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18. Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks
19. Roberto Verdecchia (VU), Architectural Technical Debt: Identification and Management
20. Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21. Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22. Sihang Qiu (TUD), Conversational Crowdsourcing
23. Hugo Manuel Proença (LIACS), Robust rules for prediction and description
24. Kaijie Zhu (TUE), On Efficient Temporal Subgraph Query Processing
25. Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26. Benno Kruit (CWI & VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27. Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28. Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022

1. Judith van Stegeren (UT), Flavor text generation for role-playing video games
2. Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
3. Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
4. Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
5. Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
6. Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
7. Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
8. Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
9. Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10. Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11. Mirjam de Haas (UvT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring