Advance Access publication on 13 June 2022
https://doi.org/10.1093/iwcomp/iwac013

Designing and Evaluating Iconic Gestures for Child-Robot Second Language Learning

Jan de Wit[1,*], Bram Willemsen[2], Mirjam de Haas[3], Rianne van den Berghe[4,5], Paul Leseman[4], Ora Oudgenoeg-Paz[4], Josje Verhagen[6], Paul Vogt[7] and Emiel Krahmer[1]

[1]*Department of Communication and Cognition, Tilburg University, Tilburg, the Netherlands*

[2]*Department of Intelligent Systems, KTH Royal Institute of Technology, Stockholm, Sweden*

[3]*Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, the Netherlands*

[4]*Department of Development of Youth and Education in Diverse Societies, Utrecht University, Utrecht, the Netherlands*

[5]*Section Leadership in Education and Development, Windesheim University of Applied Sciences, Almere, the Netherlands*

[6]*Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, the Netherlands*

[7]*School of Communication, Media & IT, Hanze University of Applied Sciences, Groningen, the Netherlands*

Corresponding author: j.m.s.dewit@tilburguniversity.edu

In this paper, we examine the process of designing robot-performed iconic hand gestures in the context of a long-term study into second language tutoring with children of approximately 5 years old. We explore four factors that may relate to their efficacy in supporting second language tutoring: the age of participating children; differences between gestures for various semantic categories, e.g. measurement words, such as *small*, versus counting words, such as *five*; the quality (comprehensibility) of the robot's gestures; and spontaneous reenactment or imitation of the gestures. Age was found to relate to children's learning outcomes, with older children benefiting more from the robot's iconic gestures than younger children, particularly for measurement words. We found no conclusive evidence that the quality of the gestures or spontaneous reenactment of said gestures related to learning outcomes. We further propose several improvements to the process of designing and implementing a robot's iconic gesture repertoire.

RESEARCH HIGHLIGHTS

- Based on existing literature, we have identified four factors that may relate to the effectiveness of robot-performed iconic gestures in supporting second language tutoring: gesture comprehensibility, the age of the learner, differences between semantic categories and spontaneous gesture reenactment.
- In the current study we found that older children (within the range of 5–6 years old) appeared to benefit from the robot's gestures, and this was particularly the case for specific semantic categories (e.g. measurement words).
- More research is needed to further study the relationship between these four factors and the effectiveness of robot-performed iconic gestures, and to investigate potential interplay between these factors.

Keywords: nonverbal communication; human–robot interaction; social robotics; second language learning

Handling Editor: Mr. Gavin Sim

Received 15 September 2021; Revised 8 March 2022; Accepted 29 March 2022

1. INTRODUCTION

There is an increasing interest in the use of robots for educational purposes (Belpaeme *et al.*, 2018, Mubin *et al.*, 2013, Toh *et al.* 2016). They can be used as a subject of learning, for example by building and programming robots together with students to teach them about robotics, artificial intelligence or computer programming. Alternatively, *social* robots can take on the role of tutors by presenting educational content and engaging in teaching activities in a multitude of domains (Mubin *et al.*, 2013), including language learning, which is the focus of the current study. One of the main motivations that drive the use of technology, and social robots specifically, in education is the need to accommodate individual children's diverse needs while at the same time the average number of students per teacher is increasing (Blatchford & Russell, 2020). By working alongside teachers (and certainly not replacing them), robots can present a cost-effective way of expanding and personalizing the content that can be offered to learners. On top of the functional role of presenting educational content, which can also be done with other tools such as tablets, robots are arguably able to provide additional social support, for example by providing (non-verbal) feedback and giving empathic responses rather than focusing merely on knowledge transfer, which has been shown to enhance the learner's engagement, as well as learning outcomes (Saerbeck *et al.*, 2010).

An important part of a robot's perceived social intelligence (Fong *et al.*, 2003) is its ability to use non-verbal communication such as gestures. Pointing (*deictic*) gestures, for instance, can be used to guide the attention of the learner toward the educational content, by referring to relevant objects (Sauppé & Mutlu, 2014). *Iconic* gestures, which are closely related in shape or motion to the concept being described (McNeill, 1992), can be used to ground new knowledge in familiar concepts or actions from the real world (Barsalou, 2008). For example, a *ball* can be depicted by molding a sphere with one's hands (shape), or by kicking an imaginary ball (motion). One particular domain that appears to benefit from gestures is (second) language learning (Hald *et al.*, 2016, Rohlfing, 2019), a domain that has recently also gained considerable attention from research into educational robots [see, e.g. Kanero *et al.*, 2018b, van den Berghe *et al.*, 2019, for reviews of existing work]. In second language learning, gestures can be used as a bridge between unknown words in the second language and existing knowledge of concepts or experiences (Hald *et al.*, 2016). In other words, gestures can be used to link the learner's non-linguistic (e.g. motor, visual) knowledge of a concept to the linguistic form of said concept.

However, robot-performed gestures may have to look different from what we are used to, as current commercially available robots are more limited in their motor degrees of freedom than humans. For example, most commonly used robots are not able to move individual fingers, making it hard to perform finger-counting or detailed hand gestures (Vogt *et al.*, 2017). This means that, in many cases, it is not possible to directly copy the way humans perform a gesture onto a robot, potentially resulting in a loss of information, which reduces the communicative ability of the gesture. This raises the question whether robot-performed gestures are able to provide the same beneficial effects to learning that we see with human-performed gestures (e.g. Hostetter, 2011, Roth, 2001).

In a previous study we investigated whether a NAO humanoid robot could support its second language tutoring efforts with iconic gestures. We found that children of 4–6 years old retained more words over time and were more engaged during the interaction if a robot used iconic gestures when introducing words in the second language, as opposed to a robot that did not use such gestures (de Wit *et al.*, 2018). In a follow-up to this previous work, which will provide the basis for the current paper, we have made several adjustments to the set-up of the study: instead of the highly iconic animal names that were taught in the previous study, the follow-up included concepts for which it is more challenging to come up with gestures with a high degree of iconicity, such as prepositions (*next to*) and comparatives (*most*). Additionally, the follow-up study consisted of seven sessions with the robot, instead of the single session in the first study. The follow-up study was conducted with children of a similar age group as the previous study, and it included a larger sample. In this case, however, we found no effect of the robot's use of iconic gestures on children's learning outcomes (Vogt *et al.*, 2019).

These mixed findings across the two studies, combined with the overall positive results found in literature on both human-performed and robot-performed gestures in supporting language learning [e.g. Hald *et al.*, 2016, van Dijk *et al.*, 2013), show us that it is important to carefully consider the design and implementation of the robot's gestures, and to investigate any contextual factors that may have prevented children from benefiting from them in our second study.

Based on existing studies into human gesturing, we have identified four factors that may relate to the effectiveness of robot-performed gestures in the context of education. First, iconic gestures only appear to contribute to learning if their meaning is clear and congruent with what is conveyed via speech (Kelly *et al.*, 2009, Macedonia *et al.*, 2011). It is therefore important that the gestures are designed in such a way that they are comprehensible for the learner. Second, the ability to interpret the meaning of iconic gestures develops during a child's early years (Novack *et al.*, 2015, Stanfield *et al.*, 2014) and, as such, the children in our study were generally at an age (5–6 years old) where they should be able to interpret the gestures. However, the fact that they were performed by a robot, with certain physical limitations and a different morphology from humans, might have negatively affected this ability. Age could therefore have played a role in the effectiveness of the robot's gestures. Third, studies have shown indications that gestures may have a greater contribution to teaching the linguistic forms of certain types of concepts

(e.g. motor events such as *running*), compared to others (e.g. de Nooijer *et al.*, 2013, Hostetter, 2011). Finally, research with human-performed gestures in the context of language learning suggests that reenactment or imitation of the teacher's gestures by the learner could further strengthen their contribution to the learning process (Repetto *et al.*, 2017, Tellier, 2005). Based on the outcomes of our previous studies, combined with the four factors identified from existing literature, we pose the following research question:

(RQ) To what extent do (1) the comprehensibility of the robot's gestures, (2) the age of participating children, (3) different semantic categories and (4) gesture reenactment relate to the successful application of robot-performed iconic gestures in second language tutoring for children?

In the current paper, we build upon our previous study (Vogt *et al.*, 2019). This is done, firstly, by thoroughly reflecting upon and conducting an evaluation study on the design of the robot's gestures, in order to find ways to improve the gesture design process. Secondly, we provide additional analyses of the data that were previously collected, focusing specifically on the four aforementioned contextual factors: comprehensibility, age, concept-based differences and reenactment. Our aim with this work is to present concrete guidelines for the design and implementation of iconic gestures for social robots, in order to optimally make use of the beneficial effects that the robot's gestures could have on (second) language learning. In the following sections we provide an overview of existing research in the field of robots for education and gestures and we cover previous work that investigated gestures performed by robots, particularly focusing on studies in education. We then introduce the set-up of the experimental study that was conducted in order to investigate the effects of a robot's use of iconic gestures to support second language learning, from which the data are used in the current analyses, and describe in detail the process of designing the robot's iconic gestures. Finally, we present and then discuss the results of our evaluation of the comprehensibility of the robot-performed iconic gestures, as well as the role of age, item-based differences and reenactment.

2. BACKGROUND

2.1. Social robots in education

The potential use of social, humanoid robots in education has become a recent focus of attention in research and in society. Next to their functional goal of presenting educational content, robots are also able to fulfill a social role that is conducive to learning, because people tend to assign human-like characteristics to them (Duffy, 2003), and therefore want to communicate with them in a human-like way (Bartneck & Forlizzi, 2004). This enables robots to teach meta-cognitive skills such as thinking aloud that can further support learning (Ramachandran *et al.*, 2018). A socially intelligent robot (Fong *et al.*, 2003) is able to observe the emotions of others and adjust its behavior accordingly (Gordon *et al.*, 2016, Szafir & Mutlu, 2012), and it can also display emotions of its own, thus showing a certain personality or character (Breazeal, 2004, Robert *et al.*, 2020). Furthermore, it is able to engage in a dialogue with human interlocutors using natural language and support its communication with non-verbal behavior such as gaze and gestures (Anzalone *et al.*, 2015, Scassellati, 2002). Its socially intelligent behavior enables the robot to build rapport, which in turn elicits more social behavior, such as constructive help-seeking (Howley *et al.*, 2014), from the learner as well. The bond between robot and learner can be further strengthened by personalizing the interactions, for example by addressing learners by their names and engaging in small talk by asking them about their interests. This can stimulate others to open up and engage more with the robot (Henkemans *et al.*, 2013). However, research by Kennedy *et al.* (2015) shows that caution is advised when designing educational human-robot interactions, as it is also possible for a robot to become *too* social, which could have a detrimental effect on learning.

Compared to virtual agents that can offer similar advantages in education, robots additionally have a physical presence in the context of the learner, which is suggested to stimulate social behavior and result in greater learning gains (Belpaeme *et al.*, 2018). A robot that is physically present is also generally rated more positively and regarded as more persuasive than a telepresent robot that is displayed on a screen or a virtual agent (Li, 2015). Furthermore, people are more likely to comply with tasks that can be seen as unusual (e.g. putting books in the trash), which rely heavily on trust (Bainbridge *et al.*, 2011), when these tasks are presented by a physically present robot instead of a virtual agent. Depending on the educational domain in which the robot is active, its ability to move within and interact with the physical world could be used to support its teaching activities (Özgür *et al.*, 2017), for example by providing realistic feedback on tasks that require manipulations in the physical world, and it allows the robot to perform classroom management (Kanda *et al.*, 2012).

One particular educational domain in which robots are commonly deployed is language learning. Because robots are seen as socially present entities, it is possible to create an immersive, natural context where learners can engage in conversations with the robot in order to facilitate language learning by immediately applying newly acquired skills in practice while receiving feedback (Lee *et al.*, 2011). Chang *et al.* (2010) further highlight the robot's ability to tirelessly repeat content and the potential use of body language to support language learning. A study by Alemi *et al.* (2015) reports that children felt less anxious, were more motivated and reported higher levels of enjoyment when training second language vocabulary with a robot compared to when no robot was present.

Previous research by Han *et al.* (2008) investigated the difference between children learning a second language from

a robot, web-based instruction and a book with audiotape in the context of their homes. Content was kept similar by taking the design for the web-based instruction and turning it into static imagery for the book and by displaying it on the robot's embedded tablet screen. They found that children were more interested and focused and performed better when a robot was used. However, Kory-Westlund *et al.* (2015) compared language learning from a robot, tablet and human teacher, and did not find any differences in terms of learning outcomes, although children did indicate that they preferred learning from the robot over the tablet and human teacher. To summarize, existing research shows promising results regarding the use of social robots in education. Their physical embodiment and presence in the context of learning set robots apart from other educational tools, such as tablet devices. Gestures could form an important way to make use of the robot's physical presence.

2.2. Gestures in education

Gestures are generally defined as 'visible actions' portrayed with our bodies (Kendon, 2004). The use of gestures plays an important role in our communication with others, for example by guiding the attention of listeners, and by making it easier for them to understand information that is communicated verbally (Hostetter, 2011). In communication, we use different types of gestures, including rhythmic *beat gestures* to emphasize certain parts of our speech, *deictic gestures* such as pointing to direct attention toward a specific entity and *representational* or *iconic gestures* in which the hands or body are used to depict a particular action, object or concept that may not be physically present (McNeill, 1992). The concept that an iconic gesture refers to is represented in some way by the motion itself, for example by pretending to brush our teeth when trying to describe a *toothbrush*, or by molding the shape of an imaginary *ball* in the air. In the current study the robot employed occasional deictic gestures to guide attention, but we focus mainly on investigating the use of iconic gestures.

Gestures, and iconic gestures in particular, are often used spontaneously and together with speech, although silent gesture or pantomime that act as a substitute for speech occur as well (McNeill, 1992). The use of gestures is an important tool in educational settings (Kelly *et al.*, 2008), where it can be considered a form of scaffolding that helps the learner understand the materials, which is particularly useful when concepts are complex or newly introduced (Alibali & Nathan, 2007). Additionally, teachers are able to hold the students' attention for longer periods of time when they use gestures to support their teaching (Valenzeno *et al.*, 2003). Specifically in second language learning, gestures can serve as a bridge between a concept that is familiar to someone in their native language, referred to as L1, and its still unknown translation in the second language (L2) by grounding the new L2 word in existing knowledge of actions or objects (Barsalou, 2008).

2.2.1. *Meaningful, comprehensible iconic gestures*

Several studies have examined the added value of iconic gestures for (second) language learning [see, e.g. Hald *et al.*, 2016, Rohlfing, 2019, for a review]. For example, Kelly *et al.* (2009) compared between L2 word learning without support from gestures, without gestures but with repeated speech, with congruent gestures or with incongruent gestures (which were the same gestures as in the congruent condition, but produced with other words than to which they belonged). Participants who received support from congruent gestures learned most words, followed by the group that received repeated speech input, the group without any additional cues, and lastly the group that received incongruent gestures. Macedonia *et al.* (2011) conducted a study in which they compared between the use of iconic gestures and meaningless gestures to support learning of an artificial language, and they found that using iconic gestures resulted in better learning outcomes than when meaningless gestures were used. Both studies show that the role of iconic gestures goes beyond merely drawing attention to the speaker, and that it is relevant to design gestures in such a way that they communicate the right meaning. Based on this, we pose the following subquestion to guide our research:

(Q1) How does the comprehensibility of the robot's iconic gestures relate to learning?

2.2.2. *Age of the learner*

We learn to interpret iconic gestures at a relatively young age. Novack *et al.* (2015) compared between 2- and 3-year-old children, and found that 2-year-olds could already take advantage of iconic gestures in the context of learning how to use new toys, although not as much as 3-year-olds. Another study by Stanfield *et al.* (2014) found that children start to understand non-redundant iconic gestures (e.g. a combination of 'read' in speech with an iconic gestures for *book*) by the age of 3 years, and that this skill continues to develop as they grow older. Existing research highlights a number of additional factors that may influence the effects of iconic gestures on communication and learning. For example, children with weaker L1 skills generally benefit more from gestures than people that have stronger L1 skills (Rowe *et al.*, 2013). Children were found to especially find support in gestures when the spoken part of the message was complex (McNeil *et al.*, 2000), potentially also due to their still developing language skills. These individual differences, particularly at a younger age as our ability to interpret gestures is still developing, leads us to the second subquestion:

(Q2) What is the role of age in the effects of the robot's iconic gestures on learning?

2.2.3. *Concept-based differences*

It is further suggested that the positive effects of iconic gestures are stronger when they describe spatial concepts (e.g. spatial relations such as *under*) or motor events (e.g. actions such as

running) than when the concepts are more abstract, such as colors, where the link between the motions and the referent is less clear (Hostetter, 2011). However, a study by Repetto *et al.* (2017), in which young adults were taught a number of abstract words (e.g. *boredom* and *alternative*) in an artificial language, still showed that participants remembered more words when they were presented to them in combination with gestures, than when the words were presented with pictures or with no additional cues. Research further suggests that verbs are especially challenging for children to learn, because children have difficulty generalizing from the particular objects or context with which they were originally taught. Because gestures do not involve interactions with real physical objects, they support the acquisition of generalizable verb knowledge better than actually performing the action on a specific tangible object (Wakefield *et al.*, 2018). In summary, research on potential differences in the effectiveness of gestures based on concept or word types is scarce, but provides a first indication that such differences do exist. As a result, we pose the following subquestion for the current research:

(Q3) Are there (item-based) differences in the contribution of gestures in supporting learning, depending on the types of concepts that are depicted?

2.2.4. *Gesture reenactment*

One important aspect of the study by Repetto *et al.* (2017), which might support learning by means of gestures, is that participants were asked to reenact or imitate the movements after observing them on screen, rather than merely observing them. In a study by Cook *et al.* (2008) children of 8–10 years old were asked to mimic the instructor's behavior when solving mathematical problems, which led to better long-term retention of the instructions compared to children that did not perform gestures themselves. Tellier (2005) found similar effects, first in the context of L1 vocabulary learning, where 42 children (5–6 years old) were split into three groups: one group was asked to repeat the words, the second also repeated the words and observed corresponding gestures, while the third group repeated words and imitated the gestures. The group that mimicked gestures performed significantly better in a short-term recall test than both other groups.

In a follow-up study (Tellier, 2008), 20 children within the same age group as the previous study learned L2 vocabulary over the course of multiple sessions. They either received pictures of the concepts that the words related to as support, or video recordings of people performing gestures for these concepts. If they were shown gestures, the children were asked to imitate them. The group of children who encoded the words using gestures performed better on the assessments, particularly on tests of their active knowledge (production, rather than recognition of the L2 words), than the group who observed pictures (Tellier, 2008). In a study by de Nooijer *et al.* (2013), children of 9–11 years old learned L1 verbs and were divided into four groups. One group only observed corresponding gestures while training the words, while the other three groups imitated the gestures, either 2) during training, 3) while trying to recall the verbs on the post-test or 4) in both situations. The results of this study indicated that imitation was only helpful for the object-manipulation verbs that were present in the study, and not for the locomotion or abstract verbs.

These findings regarding the potential benefits of enacting in order to memorize concepts align with the notion of embodied cognition, and the language-action connection (Glenberg & Gallese, 2012, Hostetter & Alibali, 2008). Although, to our knowledge, there is no existing research that draws a direct comparison between observing and reenacting iconic gestures in the context of L2 learning, based on findings in other educational domains and L1 learning we expect that children who (spontaneously) reenacted gestures in the current study may have benefited more from them than those who did not reenact, therefore we pose the following subquestion:

(Q4) Does reenactment (mimicry, imitation) of the robot's iconic gestures by the learners improve learning outcomes?

To summarize, iconic gestures have proven to be valuable tools to support education, particularly in the domain of second language learning. Their contribution to learning appears to be dependent on several factors, including the characteristics of the learner, the materials that are being taught, and whether the gestures are merely observed or also imitated. We aim to investigate whether these same factors play a role in human-robot interaction.

2.3. Related work on robots and gestures

Because robots are generally more limited in their motor degrees of freedom, their gesturing capabilities are not as extensive as that of humans, or modern virtual agents that are driven by motion capture recordings. This raises the question whether robots are expressive enough to be able to leverage the aforementioned benefits that gestures provide in human-human communication, specifically in educational contexts. Bremner & Leonards (2016) compared between co-speech iconic gestures produced by a human, and the same gestures copied to a NAO robot, the same robot used in the current study, using motion capture techniques. They found that for most gestures the adult participants in their study were able to identify the meaning in a multiple choice task equally well, regardless of whether they were performed by a human or a tele-operated robot.

Not only do the robot's gestures appear to support its communicative efforts, robots that add a non-verbal component to their speech output are also perceived differently from those that do not. A study by Salem *et al.* (2013) found that a

robot that used gestures was perceived as more human-like and likeable than one that did not use gestures, even more so when the robot made errors by performing motions that were incongruent with its speech, although at the cost of task performance. Gestures can also be used to give a certain personality or emotional state to the robot, which in turn could lead to richer, more personal interactions and to further improve people's attitude toward the robot (Aly & Tapus, 2013, Craenen *et al.*, 2018). Furthermore, several studies have reported higher levels of engagement when robots use gestures, compared to when they are static or perform random movements (de Wit *et al.*, 2018, Bremner *et al.*, 2011). In a review by Li (2015), the results from several studies indicate that people's attitude tends to be more positive toward a physically present robot compared to one that is telepresent (i.e. displayed on a screen) and to virtual agents, but only when it is using gestures—the opposite effect was found when the robot did not use gestures. This is an indication that one of the main advantages of a robot that is physically present over virtual alternatives is that it is able to move and communicate in the real world context.

Ahmad *et al.* (2016) conducted an interview study with primary and high school teachers. The teachers agreed that social robots could be useful for language learning, and they stressed the importance of gestures in language education (with and without robots). Empirical research specifically into the effects of a robot's use of iconic gestures in the context of (second) language learning is however still scarce. In a study from the related field of information retention, van Dijk *et al.* (2013) showed in a single session with adult participants that the use of iconic gestures by a robot increased retention, particularly of verbs, measured using a recall task. Similar results on information retention were found in the context of storytelling (Bremner *et al.*, 2011, Huang & Mutlu, 2013). Another study involving storytelling by a robot further suggests that exaggerated gestures, which are perceived as more cartoon-like, lead to increased memorization of the story compared to 'normal' (unexaggerated) motion, and the robot was perceived as more engaging and entertaining when exaggerating its movements (Gielniak & Thomaz, 2012).

In previous work with 4- to 6-year-old children, we have shown that the robot's use of iconic gestures while presenting words (animal names) in a second language aided the recall of these words approximately 1 week after training, and resulted in an overall higher level of engagement of the child while learning with the robot (de Wit *et al.*, 2018). Although these gestures were intentionally chosen and designed to have a high degree of iconicity, the results of this study do serve as a first indication that the benefits of iconic gestures that we see in human–human tutoring situations could apply to robot-performed gestures as well. After this initial exploration, we conducted a large-scale study to further investigate the potential application of social robots in second language tutoring. In Vogt *et al.* (2019), we have concisely described the learning effects in the different conditions (briefly summarized in the next section), which provides the basis for the current paper. In this paper, we present an in-depth analysis of the design and the effects of the robot's use of iconic gestures, which was not part of Vogt *et al.* (2019).

2.4. Large-scale study

The previous study, from which data are further analyzed in the current paper, was conducted at nine different primary schools throughout the Netherlands, in which children of approximately 5 years old ($M = 5$ years, 8 months; $SD = 5$ months) interacted with an intelligent tutoring system (ITS). This ITS consisted of a tablet device on which educational content was shown, and a robot that engaged in learning activities with the children. The study included seven sessions, where new L2 vocabulary was introduced in the first six, while the seventh session served as a recap of the previously taught words. Our aim was to investigate the following: (1) whether the ITS is effective at teaching children L2 vocabulary, (2) whether the robot's physical presence contributes to learning outcomes (compared to only using a tablet) and (3) whether robot-performed iconic gestures result in greater learning outcomes, compared to a robot that does not use iconic gestures. In order to study these effects, we assigned the children to one of the following conditions:

(i) **Control (no treatment)**, where children had an interaction with the robot once a week (for a total of three interactions), which did not involve any educational content related to second language vocabulary. This was to control for the possibility that children were exposed to the target vocabulary outside of the language learning lessons.
(ii) **Tablet only**, where children interacted only with the tablet. The robot was hidden from view, with its speech output routed through the tablet's speakers.
(iii) **Tablet + robot without iconic gestures**, where children interacted with the tablet and the robot, and the robot would use deictic (pointing, tablet manipulation) gestures to guide the child's attention.
(iv) **Tablet + robot with iconic gestures**, where children interacted with the tablet and the robot, and the robot would use both deictic gestures to guide the child's attention and manipulate objects on the tablet, as well as a corresponding iconic gesture whenever it pronounced one of the target words in the second language.

A total of 194 children, 97 boys and 97 girls, participated and met the inclusion criteria (e.g. scoring a maximum of 17 out of 34 words correct on the English translation pre-test). They were pseudo-randomly assigned to the experimental conditions, with a balance in age and gender, resulting in 32 participants in the control condition (i), and 54 participants in each of the three experimental conditions (ii–iv). The children's legal guardians gave informed consent, and the study was approved by our

institutions' research ethics committees. The study and analysis plan were preregistered on AsPredicted[1] .

The results, which are presented in detail in Vogt *et al.* (2019), showed that children in the three experimental conditions scored significantly higher on translation as well as comprehension tasks, than those in the control condition (all P-values < 0.01). This means that the tutoring interaction was effective. However, contrary to our expectations, no significant differences were found between the three experimental conditions of tablet only, tablet + robot without iconic gestures and tablet + robot with iconic gestures. In other words, there was no observed effect of the robot's physical presence and use of deictic gestures, nor of its use of iconic gestures, on the students' learning outcomes. For the remainder of this paper, we will focus our attention on the robot's use of iconic gestures, to get a better understanding of the role of these gestures in the child-robot interactions.

It is important to note that we take an exploratory, inductive approach in this paper. The research questions for the current paper were formulated after conducting our main, preregistered analyses (as described in Vogt *et al.*, 2019), and therefore the experiment could not be designed in such a way that the four factors currently under investigation could be empirically tested. Our intention now is to further contextualize the main results of the experiment, particularly those pertaining to the robot's use of iconic gestures, and to propose an agenda for future robot gesture research in the field of education, based on theories from human gesture studies that were outlined in the previous sections.

In the following section, we first describe the design of the ITS as a whole. This is important, because the iconic gestures were included as part of this tutoring interaction and were not used in isolation, therefore the nature of this interaction (and how it is different from other studies) could potentially have influenced the effectiveness of the robot's iconic gestures. We then introduce the process of designing the gestures, and what the resulting gestures looked like. The measurement instruments are then presented, as these are needed to interpret the analyses that follow. Finally, we present the results of our analyses, and conclude with a discussion of our findings and recommendations for the design and implementation of robot-performed iconic gestures.

3. INTERACTION AND GESTURE DESIGN

3.1. Design of the tutoring interaction

The ITS consisted of a Softbank Robotics NAO V5 robot, combined with a Microsoft Surface Pro 4 tablet through which the child engaged in the learning interaction. The robot was placed in a crouching position at a 90-degree angle relative to the child. This helped to position the robot as a peer rather than a teacher, which has been shown to result in increased task engagement and performance (Zaga *et al.*, 2015). In addition, this made it easier for the learner to take on the robot's perspective, thereby avoiding confusion for gestures such as *left*, which would be harder to interpret if the robot would be sitting directly across from the learner. Figure 1 shows the general positioning of the robot and tablet. This positioning was kept as consistently as possible between different schools. One camera was placed facing the child, with a second camera to the side and behind the child, so that the interactions with the tablet could also be recorded. To make the robot seem more life-like, we enabled 'breathing mode', which caused its arms to move around slightly, giving the illusion that the robot was actively breathing. It also blinked its eyes every few seconds and was tracking the child's face to establish eye contact.

FIGURE 1. Positioning of the tablet and the robot during the experiment.

The content of the study comprised seven lessons in total. The first six lessons each took place in a different virtual environment, such as a forest or a playground, where the native Dutch-speaking child was introduced to five or six words in the second language (English) during each lesson (see Figure 2 for an example). We opted for the use of virtual environments and objects instead of physical ones because automatic perception and manipulation of real objects in a dynamic physical context would have been challenging to implement. Virtual objects have also been shown to be equally effective in supporting math and L1 teaching (Klahr *et al.*, 2007, Singer & Gerrits, 2015). Moreover, in a preliminary study comparing the effects of physical versus virtual objects on L2 learning, we did not find differences in learning outcomes (Vlaar *et al.*, 2017).

In the first three lessons, the target words belonged to the number domain, including concepts such as counting words (*one, two, three, four, five*), mathematical operations (*add, take away*) and comparisons (*more, most*). Lessons four, five and six focused on spatial relations and verbs, which contained words such as *above, next to, walking* and *sliding*. These words were selected based on a survey of existing educational

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

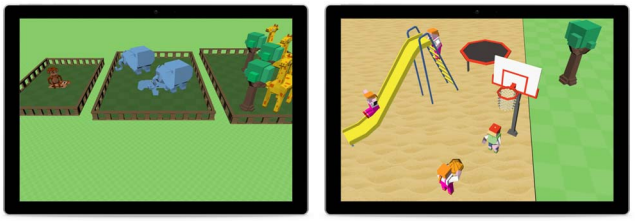

FIGURE 2. Examples of the virtual environments shown on the tablet (left: lesson one in the zoo, where animals have been brought back to their cages; right: lesson six in the playground, which was first 'built' by placing equipment, and now children started playing in the area).

TABLE 1. English words included in the study, per lesson.

Lesson	Environment	English words
1	Zoo	One, two, three, add, more, most
2	Bakery	Four, five, take away, fewer, fewest
3	Zoo	Big, small, heavy, light, high, low
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
7	Photo book	Recapitulation of all words

curricula, word frequency and age of acquisition lists[2] to ensure that children were familiar with the concepts in their native language. The final seventh lesson did not introduce any new target words, but instead recapitulated all 34 target words from the previous six lessons. Table 1 shows a list of all the English words that were included in the study, as well as the virtual environment in which they were presented. The children were not encouraged nor discouraged to practice the English words outside of these seven lessons. Parents, caregivers and teachers were unaware of the exact target words that were taught, to avoid external influences as much as possible.

During each lesson, children went through a particular scenario together with the robot, while they completed several different tasks that were presented to them by the robot, such as touching or moving objects on the screen, repeating words after the robot or performing an action in the real world (such as pretending how to climb). To further position the robot as a peer, the tablet was used to actually initiate these tasks, for example by making new objects appear on the scene. The robot would then observe this change on the tablet and suggest the course of action in order to continue, as if the robot and child were learning together, for example by stating that 'the monkey has escaped — let's put it back in its cage!'.

[2] https://web.archive.org/web/20210415022714/http://www.l2tor.eu/effe/wp-content/uploads/2015/12/D1.1-Lessons-series-three-domains.pdf

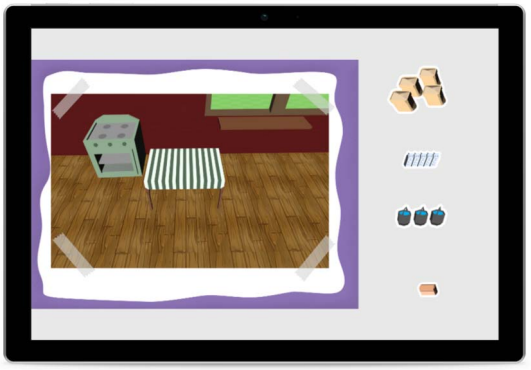

FIGURE 3. The photo book environment used in the seventh (recapitulation) lesson.

The lessons followed predefined scripts, so that each child experienced the same interaction by performing the tasks in the same order. The scripts were created in such a way that all target words were mentioned at least 10 times during the lesson in which they were introduced, and once more in the lesson that followed it. In the condition with iconic gestures, the robot would perform the corresponding gesture whenever it pronounced a target word in the L2. If the child did not perform any action or if the action was incorrect, the robot would repeat the task up to two times, which resulted in additional exposures to the English words and, in the condition with iconic gestures, the corresponding gestures. If the task was still unfinished after two reminders, the robot performed the task for the child, for example by moving objects on the screen or by counting down and then repeating the words together with the child, to ensure that the script was always completed. Because of the robot's imperfect pronunciation, the first mention of each word was by means of a recording from a native English speaker, which was played back through the tablet, generally as a response to the child successfully performing an action such as touching or moving an object on the screen.

During the seventh lesson, which was a recapitulation of the previously learned words, children constructed a photo book that contained six pages, each with a screenshot of the backdrop of one of the previous lessons. The children were then asked to drag stickers containing objects that were present in the original scenes onto the pages, while practicing the related English words. Figure 3 shows one of the pages of the photo book, with the stickers not yet placed. While the other lessons all had three-dimensional environments, the recapitulation lesson was two-dimensional. Although all 34 target words had to be covered in this lesson, there were fewer repetitions of these words compared to previous lessons, resulting in a total session length of approximately 15–20 minutes, which was similar to the other six sessions.

The researcher had a control panel running on a laptop, which could be used to start a specific lesson. This was also used to enter the child's name, so that the robot could use it

during the interaction, and it provided the researcher with the option to pause the lesson if needed. The robot acted nearly autonomously, with the exception of recognizing whether children successfully completed tasks in which they had to repeat words after the robot, or had to enact a certain action, for which the sensing techniques were difficult to implement. For example, the use of automatic speech recognition (ASR) to detect whether children correctly repeated after the robot is not yet reliable enough (Mubin *et al.*, 2012), especially when attempting to recognize young children's speech (Kennedy *et al.*, 2017). For tasks where the child had to repeat a target word, the researcher therefore pressed a button on the control panel when the child spoke for the interaction to continue (a Wizard of Oz approach). For other points during the scenario where we expected a reply from children (e.g. during small talk, or in the case of enactment), we implemented pauses to create the illusion that the robot was watching and listening to the children. To give an impression of what the interaction between child and robot looked like, we refer to a promotional video that was developed as part of the L2TOR project[3] .

3.2. Design of the robot's gestures

3.2.1. *Deictic gestures*

The robot performed three types of deictic gestures during the tutoring interactions with the children (Figure 4). The first type was implemented at predefined locations within the script, where the robot would point toward the tablet screen to direct the child's attention to it. This gesture was always the same, so there was no distinction between different parts of the screen—the robot directed its gaze toward the tablet, and pointed at its general direction. The other two types of deictic gestures were used when the robot provided help to the learner after a task was performed incorrectly or not performed at all. If the task was to move an object to a different location, the robot would 'swipe' across the screen while at the same time the object would move to its correct target location. A similar motion was implemented to simulate the robot touching an object on the screen. In this case the robot would extend its arm over the tablet and then briefly open and close its hand. At the same time, the corresponding object was highlighted on the screen to simulate the robot's triggering of the object. Both the swiping and touching gestures, just like the pointing gesture, were always the same and were not linked to any exact locations on the tablet. However, this proved to be realistic enough to provide the illusion of the robot performing manipulations within the virtual environment. We also explained to children that this was how the robot controlled the tablet, and this explanation was accepted by them.

3.2.2. *Designing human-like iconic gestures*

The iconic gestures for the chosen target words were based on a dataset that was collected using a gesture elicitation procedure (Kanero *et al.*, 2018a). In this elicitation study, three participants, all native speakers of English, were recorded while performing corresponding gestures for all 34 concepts. Twenty other participants, also native English speakers, were then asked to view these recordings and rate on a scale from 1 to 7 the comprehensibility of the human-performed versions of the gestures, or the degree to which they matched the words they intended to describe. Because participants in this study were not constrained to the robot's physical limitations, several gestures contained certain features or motor skills that are not supported by the NAO robot (e.g. jumping up and down or finger-counting), preventing a direct mapping from these recorded gestures onto the robot. For this reason, several gestures had to be reinterpreted, although the suggestions from the elicitation procedure were still used as a guideline. Figure 5 (left) displays an example of finger counting where such a reinterpretation had to take place: To depict the concept *four* using the robot's fingers, we had the robot raise both hands showing two of its three fingers per hand by turning the wrist so that the thumb was hidden from view. Figure 5 (right) shows a gesture that could be translated more directly, without adjustments. The gestures for the robot were made using the Choregraphe tool that is provided with the NAO robot (Pot *et al.*, 2009), which allows the designer to define keyframes. The robot then interpolates between these keyframes when producing a gesture.

An initial pilot evaluation with five verbs (out of the 34 target words) was conducted to validate whether the gestures' comprehensibility, or how well the gestures matched the concepts they intended to describe, indeed influenced how well these gestures support tutoring by leading to improved learning outcomes. This was done by conducting a between-subjects study with children as participants ($N = 43$, $M_{age} = 5$ years, 9 months, $SD_{age} = 7$ months), where the gestures were either performed by a robot or by a human tutor. The results, described in more detail in one of our project's deliverables[4] , indicate that indeed the comprehensibility of an iconic gesture, as originally rated for human-performed versions by 20 adult participants, had an effect on learning outcomes of children that used these gestures to learn English words, at least when this was measured by means of a receptive vocabulary task. No significant differences were found in a production task.

Before including them in the current experiment, the gestures were revised once more, especially taking into account the change in the robot's positioning relative to the child—in the original recordings, participants were standing and facing the camera, while in the experiment the robot was seated and

[3] https://www.youtube.com/watch?v=y8W-2XgdfoI

[4] https://web.archive.org/web/20210415022714/http://www.l2tor.eu/effe/wp-content/uploads/2015/12/D7.4-Evaluation-report-storytelling-domain.pdf

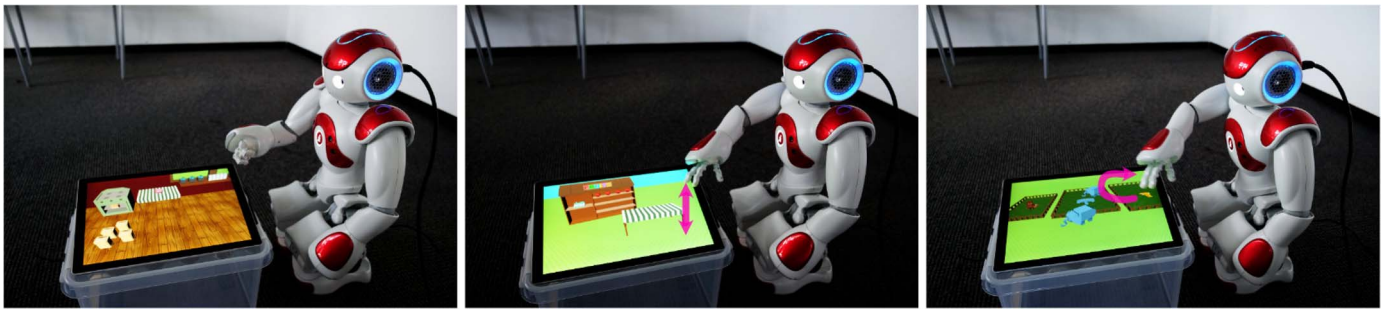

FIGURE 4. The three types of deictic gestures used in the study (left: pointing (closed hand); middle: pretending to touch the screen (the hand briefly opens and closes); right: pretending to swipe across the screen (open hand)).

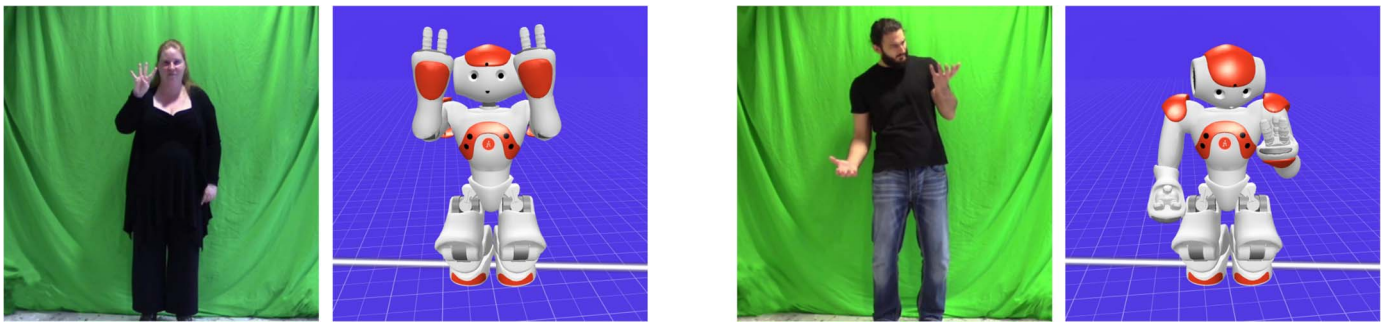

FIGURE 5. Examples of the translation of human-recorded gestures onto the robot for the concepts *four* (left) and *light* (right); images used from the data of Kanero *et al.* (2018a) with permission.

placed at a 90-degree angle to the right of the child, changing the way gestures were perceived. Figure 6 shows photographs of all 34 gestures as they were used in the study, taken from the perspective of the learner.

There is a further distinction between the gestures that were designed for this study: Examples such as *running* use the whole body, where the robot actually 'becomes' the runner (character viewpoint), while others such as *jumping* instead use one hand to depict an imaginary character or object that is jumping, also known as the observer viewpoint. Research has shown that younger children tend to use a larger gesture space, and perform gestures from the character viewpoint (as is the case with the *running* example) more often than smaller, imaginative gestures from the observer viewpoint such as the one for *jumping* (Sekine *et al.*, 2018). This suggests that it could be better to use more gestures where the robot actually 'becomes' the concept. However, this was not always possible given the robot's physical limitations.

3.2.3. *Integration with the lesson content*

The built-in text-to-speech engine of the NAO robot is able to trigger events, such as performing a gesture, at specific points during the robot's speech output. This was used to align speech and gestures, as well as perform coordinated deictic gestures and shifts in the robot's gaze to guide the learner's attention. For the iconic gestures we introduced pauses in the robot's speech, such that the corresponding target word in the L2 coincided with the stroke, the most salient part of the gesture. If possible, the pronunciation of the target word was timed for a moment with little to no movement, thereby minimizing any negative influences that motor noise could have on the audibility of the robot's speech. The robot then resumed talking in L1 after the gesture was completed.

4. DATA COLLECTION

4.1. Procedure

4.1.1. *Group introduction*

Children were first introduced to the robot in a group setting. This was generally done with an entire classroom, including children that did not (yet) sign up to participate in the experiment, with the teacher also present. Previous research has shown that these group introductions reduce anxiety for subsequent individual interactions (Vogt *et al.*, 2017, Fridin, 2014). During the group session, the robot introduced itself as 'Robin'—a unisex name, leaving the robot's gender open to interpretation—and demonstrated some of its abilities, for example by performing several dances and by inviting the children to join in taking on a number of different poses. It also highlighted some of its limitations, for example by mentioning that it could not hear very well, thereby instructing children to speak loudly. This was done so that researchers could clearly

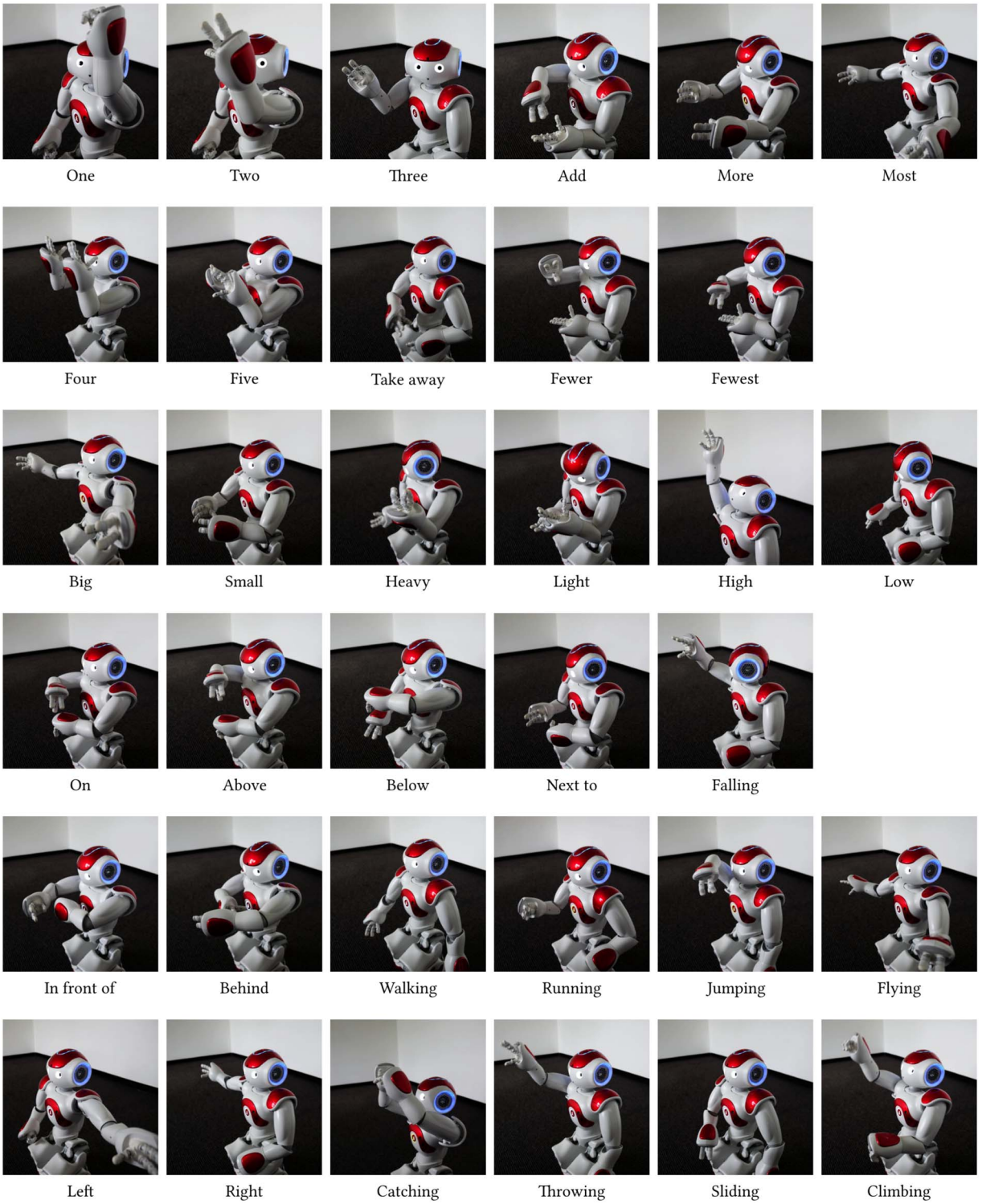

FIGURE 6. Gestures for all of the 34 concepts in the study; video recordings are available at https://youtube.com/playlist?list=PLJreGGDWkgkqQUIsZXMgekMHP1T-_dfbU.

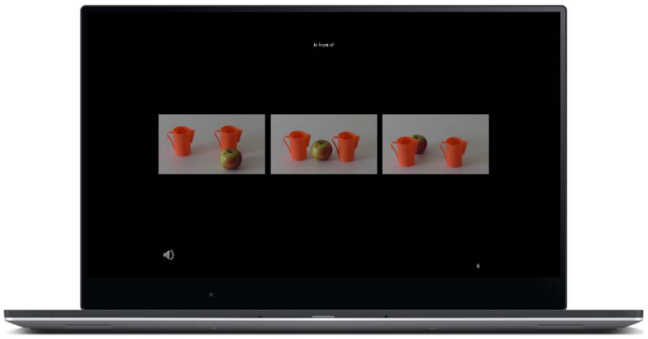

FIGURE 7. Example of the comprehension task (for the concept *in front of*) administered as part of the post-tests.

hear the children repeating after the robot during the lessons, allowing them to press the Wizard of Oz button on the control panel. Children were invited to shake the robot's hand, which may help them bond with the robot.

4.1.2. *Pre-test*

The pre-test took place either on the same day as the group introduction, or shortly thereafter. Children were picked up from their classroom one by one and brought to a separate, quiet room—often the same room in which they later interacted with the robot. They sat down at a table on which a laptop was placed, with a researcher sitting next to them. The researcher then walked through the different pre-test segments in a predefined order:

1. Peabody Picture Vocabulary Test (L1 vocabulary knowledge);
2. Translation task of the target words from L2 to L1;
3. Visual search task (selective attention);
4. Non-word repetition task (phonological memory);
5. Questionnaire measuring anthropomorphism.

Depending on the type of task, the child either answered verbally or pointed at items on the screen, while the researcher took notes on a paper sheet or pressed corresponding buttons on the keyboard. The researcher gave positively voiced neutral feedback to the child without indicating whether the answers given were correct or not. If the child did not know an answer to one of the tests, the researcher reassured them that this was not a problem and stimulated them to proceed with the tasks. After completing all segments the child was brought back to the classroom. The pre-test took approximately 45 minutes, and was recorded with a video camera.

4.1.3. *Lessons*

Children who were assigned to one of the three experimental conditions took part in a total of seven lessons, which were scheduled so that children received two lessons per week, and never two lessons on the same day. As a result most children completed the lesson plan over the course of 4 weeks. The first lesson was planned at least one day after the pre-test.

The interactions were situated in a separate, quiet room at the school, where the robot was sitting on the floor next to the tablet. The child was collected from his or her classroom and invited to sit in front of the tablet, after which the researcher started the lesson using the control panel. While the child and robot completed the lesson together, the researcher was sitting behind the child to discourage the child from looking at him or her instead of the robot for feedback. If needed the lesson could be paused and resumed using the control panel. The end of a lesson was always marked by stars appearing and moving around on the tablet screen, after which the robot said goodbye and the child was brought back to the classroom. Each session with the robot took approximately 15–20 minutes to complete.

4.1.4. *Post-test*

The post-test was administered to each child twice, first an immediate post-test close to the last lesson (but at least one day later), and then a delayed post-test approximately 2–5 weeks after the immediate post-test. In both cases the child was picked up from the classroom and brought to a quiet room. Similar to the pre-test, the child sat down at a table with the researcher sitting next to him or her. Using a laptop, the two translation tasks and comprehension task described in Section 4.2 were completed in the following order:

1. Translation from L2 to L1;
2. Translation from L1 to L2;
3. Comprehension task;
4. Questionnaire measuring anthropomorphism (only in the immediate post-test).

The researcher noted down the answers as they were given by the child. Each post-test took approximately 30–45 minutes to complete, and was recorded with a video camera.

4.2. Measures

Three different tasks were used to measure whether children learned and remembered the target words. This included two translation tasks, one from the L2 to the L1 and one from the L1 to the L2, to measure children's ability to freely produce translations of the target words. In both tasks the researcher would repeat a predefined sentence ('Wat is [word] in het [language]?' — 'What does [word] mean in [language]?'), where the word was either in L1 or L2, and the language was either Dutch or English depending on the translation task. The pronunciation of the target words was made consistent by using recordings from a bilingual speaker of Dutch and English, which were embedded in a set of Powerpoint slides and then triggered by the researcher.

To measure children's comprehension of the target words in L2, we conducted a separate task where children were shown a set of Powerpoint slides, each slide containing three pictures or videos depicting a certain concept (Figure 7). A voice recording from a native speaker was played back every

time a new slide was shown, asking 'Waar zie je... [L2 word]' ('Where do you see... [L2 word]'), after which the child was asked to point at the corresponding picture or video. Depending on the target word, these stimuli would contain several physical objects, or a person performing a certain action. Because there is a relatively large probability that the children would guess correctly (33%), each concept was tested three times using different contexts, and shown with different distractor concepts (incorrect answers). However, because this would result in too many trials if all target words were included, we only tested 18 words, which were pseudo-randomly selected to include examples from all of the semantic categories (e.g. counting, measurement, movement verbs), and from all of the six lessons. The same 18 words were used for all children to avoid a potential effect of differences in difficulty level if words were randomly picked. Multiple versions were developed of both translation tasks and the comprehension task, in which the items were presented in a different order.

We further measured the children's receptive L1 vocabulary knowledge using the Peabody Picture Vocabulary Test (Schlichting, 2005), their phonological memory with a non-word repetition task (Chiat, 2015), and selective attention by means of a visual search task (Mulder *et al.*, 2014). In addition, we investigated the extent to which children anthropomorphized the robot by means of a questionnaire (van den Berghe *et al.*, 2021a).

5. EVALUATING THE ROBOT'S GESTURES

In this section we present a detailed examination of the different factors that may have influenced the effectiveness of the robot's iconic gestures. We will first look at the comprehensibility of the gestures (Q1), followed by the role of age (Q2), differences between semantic categories of the target words (Q3) and finally the potential benefits of not only observing but also reenacting the gestures (Q4). For each analysis, we will first introduce the approach that was taken, followed by the results.

5.1. Comprehensibility of the gestures

Studies into human-performed gestures indicate that iconic gestures should actually convey meaning, as meaningless or incongruent gestures do not appear to contribute to language learning and may in fact even have a detrimental effect (Kelly *et al.*, 2009, Macedonia *et al.*, 2011). We therefore investigated whether the meaning of the gestures included in the current study was clear by means of an online evaluation with adults, and then compared these comprehensibility scores to the learning outcomes of children in the study to see whether the comprehensibility of the gesture of a particular concept contributed to learning the English word for that concept.

5.1.1. *Analysis approach*

To investigate whether the meaning of the 34 final gestures included in the study was clear, we conducted an online evaluation study for the purposes of the current paper. In this evaluation study, participants were 17 adults, 10 female and 7 male, with an average age of 21 years and 6 months (SD = 2 years, 8 months), recruited via convenience sampling. They were shown videos of all robot gestures, recorded from the same perspective as the photographs in Figure 6, in random order. Each video was between three to eight seconds long. Participants were asked to choose the concept belonging to the gesture they were just shown from a list of six possible answers. The incorrect answers were always the other concepts from the same lesson, to measure whether the 34 gestures were iconic enough to identify them within the context of the lesson in which they were used. The answers were also randomized for each trial. Lessons two (bakery) and four (fruit shop) contained only five target words in total, therefore the words *six* and *lifting* were added to these respective lessons as additional (incorrect) answers to ensure that the chance of guessing correctly was always the same.

Along with identifying the corresponding concept (binary scores, correct or incorrect), participants were asked to rate the clarity and naturalness of the gesture, by means of two separate 5-point scales ranging from 1 (extremely unclear/unnatural) to 5 (extremely clear/natural). We then calculated the accuracy for each concept, which is the number of participants in the gesture evaluation study that correctly identified the concept divided by the total number of participants, resulting in a score from 0–1, as a measure of how comprehensible the corresponding gesture was. Although this should perhaps be considered a measure of recognition of a particular gesture within a limited set of six rather than general comprehensibility, we consider this an accurate measurement of comprehensibility because the children in the study only had to distinguish between five or six gestures per lesson as well (except for the recap lesson, but this did not introduce new gestures). In addition, in this evaluation study the gestures were presented in isolation, while the robot used them in conjunction with speech during the experiment, thereby providing additional information to help children comprehend the gestures.

Correlation analysis (Kendall's tau-b, because of the relatively small sample size) was used to test whether the accuracy (as a measure of comprehensibility of the gesture), clarity and naturalness were significantly correlated. In addition, we grouped the concepts into semantic categories, such as counting words and prepositions, based on existing language learning curricula[5]. Using paired samples t-tests, we tested whether there were significant differences between the semantic categories, in terms of the comprehensibility, clarity and naturalness of the gestures.

[5] See, e.g. https://www.gov.uk/government/collections/national-curriculum

We also explored the relationship between the comprehensibility of the gesture for a concept, and children's learning outcomes for that same concept. Learning outcomes were calculated for the 54 children that were in the experimental condition with iconic gestures, based on their performance on the translation tasks. There were two translation tasks, from the L2 (English) to the L1 (Dutch) and from the L1 to the L2. Because there was a strong correlation between the two tasks, indicating that they both measure a similar language production skill, the scores on both tasks were averaged. This means that for each concept, the score of one child could be either 0 (incorrect on both tasks), 0.5 (correct on one of the two tasks) or 1 (correct on both tasks). For this analysis, we only included the experimental condition where the robot used iconic gestures, to focus on the relationship between gesture comprehensibility and the resulting learning outcomes when these gestures were used. Children's scores on the translation tasks were averaged across all children in the condition with iconic gestures ($N = 54$), to reach an average score for that particular concept (ranging from 0–1). We then compared the scores on both post-tests (immediate and delayed) for each concept to the rated comprehensibility of the gesture for that concept using correlation analysis. Note that the comprehension task of the post-tests only tested 18 out of the 34 target words, therefore we can only analyze the relationship between comprehensibility and post-test scores for these 18 words.

5.1.2. *Results: evaluation study with adults*

Appendix A shows a full overview of the comprehensibility (accuracy) scores, and the ratings of clarity and naturalness, from the adult participants in the online evaluation study. Kendall's tau-b correlation was calculated to test the relationship between participants' accuracy in identifying the concept that was described by a gesture, referred to as comprehensibility ($M = 0.72, SD = 0.09$), and the rated clarity of the gestures ($M = 3.69, SD = 0.42$). This showed a significant medium correlation, $\tau_b = 0.37, P = 0.045$, where participants who rated the gestures as more clear also had a higher chance of matching this gesture with the correct answer. In addition, the correlation between gesture clarity and naturalness was significant, $\tau_b = 0.36, P = 0.043$, indicating that gestures that were rated as more clear were generally also rated as more natural. However, the correlation between comprehensibility and the rated naturalness of the gestures ($M = 3.48, SD = 0.35$) was not significant, $\tau_b = 0.35, P = 0.06$.

Table 2 presents a summary where we grouped the concepts by semantic categories and calculated the mean scores across all words of each category. The lowest comprehensibility scores, 0.35 and 0.51, were found for 'counting words' and 'comparatives', while 'operations' and 'movement verbs' had the highest comprehensibility scores: 0.97 and 0.88. 'Measurement words' and 'prepositions' received scores of 0.75 and 0.79. In the 'measurement words', the word *heavy* scored low (0.29) compared to the other words in that semantic category, while *light*—which has a similar gesture—was generally identified correctly (0.88). For the 'prepositions', the gesture for *on* scored especially low on comprehensibility (0.47), compared to the other gestures in the same category. Using paired samples *t*-tests, we tested whether there were significant differences between the semantic categories. The results, which are presented in full in Appendix B, show that there was a significant difference in comprehensibility between all semantic categories, except for 'measurement words' and 'prepositions' ($P = 0.30$). The clarity and naturalness ratings showed similar patterns to each other: They both differed significantly between 'counting words' and 'measurement words', 'prepositions' and 'movement verbs' (all *P*-values < 0.001), between 'comparatives' and 'measurement words', 'prepositions' and 'movement verbs' (all *P*-values ≤ 0.007), between 'operations' and 'measurement words', 'prepositions', and 'movement verbs' (all *p*-values $\leq .01$), and between 'prepositions' and 'movement verbs' (both *P*-values = 0.02).

In summary, the evaluation study of the gestures with adults showed differences in the comprehensibility (accuracy at the identifying the corresponding concepts), clarity and naturalness, both between and within the different semantic categories. Particularly 'counting words' and 'comparatives' were often not correctly identified, while 'operations' and 'movement verbs' were relatively easy to recognize. Comprehensibility correlated moderately and significantly with the rated clarity, and also moderately, albeit not significantly with the rated naturalness of the gestures. Naturalness correlated moderately and significantly with clarity.

5.1.3. *Results: Comprehensibility and children's learning outcomes*

Figure 8 shows the comprehensibility scores, collected during the rating study with adults and discussed in the previous subsection, on the horizontal axis, and the children's average scores on the translation tasks in the study on the vertical axis, for both the immediate (left) and delayed (right) post-tests. By visually inspecting these graphs, we could identify three broad 'clusters', which appear for both post-tests:

1. High scores on the translation tasks, but low comprehensibility ratings—this cluster consists mainly of 'counting words', such as *four*;
2. Medium to high scores on the translation tasks, and high comprehensibility ratings—this cluster mainly includes 'movement verbs', such as *jumping*;
3. Low scores on the translation tasks, and medium to high comprehensibility ratings—this cluster includes most of the 'comparatives' (e.g. *more*), 'operations' (e.g. *take away*), 'measurement words' (e.g. *light*) and 'prepositions' (e.g. *behind*).

Kendall's tau-b correlation shows that the correlation between the comprehensibility of the gestures, as rated

TABLE 2. Comprehensibility (0–1), clarity (1–5) and naturalness (1–5) ratings for the gestures per semantic category (SD in parentheses). These are mean scores across all words of each category. Chance level for comprehensibility is 0.17.

Semantic category	Comprehensibility	Clarity	Naturalness
Counting	0.35 (.11)	3.00 (1.30)	2.94 (1.13)
One, two, three, four, five			
Comparatives	0.51 (.23)	3.22 (1.01)	3.19 (0.92)
More, most, fewer, fewest			
Operations	0.97 (.04)	3.03 (1.29)	2.94 (1.18)
Add, take away			
Measurement	0.75 (.24)	3.92 (0.89)	3.68 (0.87)
Big, small, heavy, light, high, low			
Prepositions	0.79 (.15)	3.89 (1.10)	3.56 (0.98)
On, above, below, next to, in front of, behind, left, right			
Movement verbs	0.88 (.11)	4.11 (1.12)	3.83 (1.10)
Falling, walking, running, jumping, flying, catching, throwing, sliding, climbing			

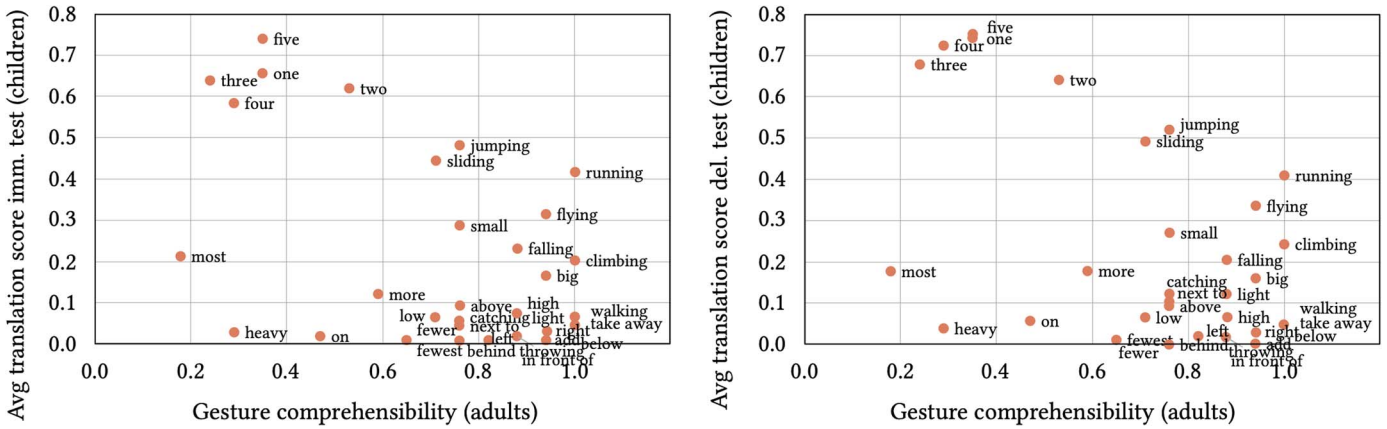

FIGURE 8. The individual gestures' comprehensibility, in terms of mean accuracy by adult raters (horizontal axis), compared to the average translation scores of children in the condition with iconic gestures ($N = 54$) for these concepts (vertical axis). Scores per child were either 0 (no correct), 0.5 (correct on 1 translation task, L1->L2 or L2->L1), or 1 (correct on both tasks). Left: immediate post-test; right: delayed post-test.

by adults, and the scores of children participating in the condition with iconic gestures on the translation tasks was negative and not significant for the immediate post-test ($\tau_b = -0.19, P = 0.13$) as well as the delayed post-test ($\tau_b = -0.20$, $P = 0.11$).

Figure 9 shows the same gesture comprehensibility scores on the horizontal axis, but now with children's average scores on the comprehension task on the vertical axis, for the immediate (left) and delayed (right) post-tests. Note that only 18 out of the 34 target words were included in this task. The results show a similar pattern for the comprehension task to the scores on the translation tasks, where children scored well on 'counting words' and 'motion verbs'. Additionally, children seemed to perform slightly better on some of the 'measurement words' (*small*, *heavy*) and 'comparatives' (*most*) on this task. Note that chance level for this score was 0.33. The Kendall's tau-b correlations between the comprehensibility ratings of the gestures, and children's performance on the comprehension task were negative and not significant for the immediate post-test ($\tau_b = -0.17, P = 0.36$), and the delayed post-test ($\tau_b = -0.17, P = 0.36$).

To summarize, we do not find conclusive evidence that there is a relationship between the comprehensibility of the gestures, as measured with adults, and performance of children in the large-scale study on the post-test tasks.

5.2. Age-based differences between learners

Based on indications in existing research that the ability to perform and interpret (iconic) gestures develops during early childhood (Novack *et al.*, 2015, Sekine *et al.*, 2018, Stanfield *et al.*, 2014), we explored how age was related to children's learning outcomes during our study, with and without the robot's use of iconic gestures.

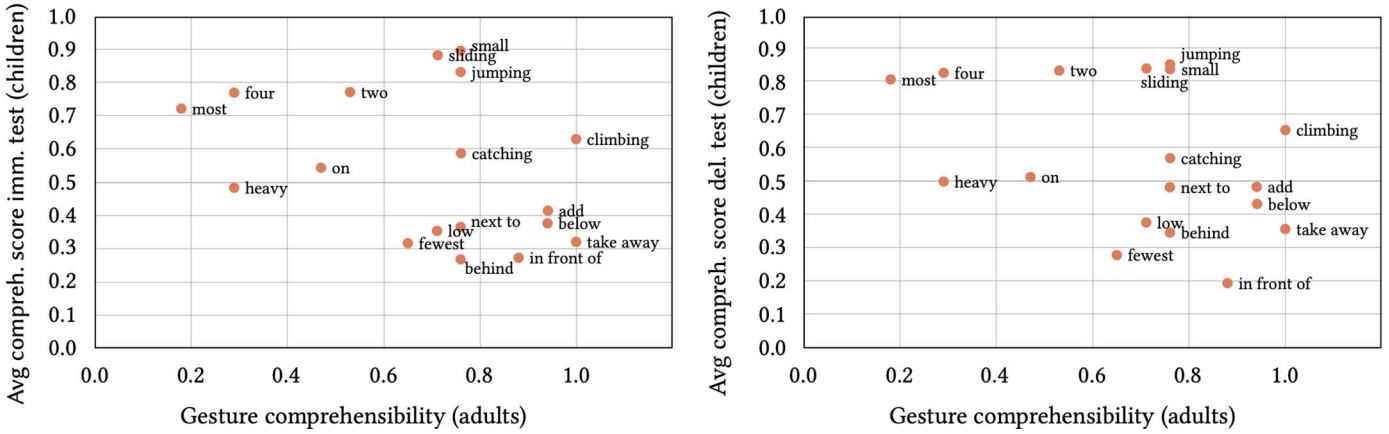

FIGURE 9. The individual gestures' comprehensibility, in terms of mean accuracy by adult raters (horizontal axis), compared to the comprehension task scores of children in the condition with iconic gestures ($N = 54$) for these concepts (vertical axis). Scores per child were either 0 (no correct), 0.33 (1 round correct), 0.66 (2 rounds correct), or 1 (all 3 rounds correct). Chance level is 0.33. Left: immediate post-test; right: delayed post-test.

5.2.1. *Analysis approach*

To study the effect of the participating children's age on their learning outcomes, we ran the same analysis that was used in the original study (Vogt *et al.*, 2019) to measure learning outcomes, but now with children's age at the time of the pre-test (in months) as a covariate. This analysis included all four conditions so that we can investigate whether an observed effect of age applies to learning in general, or only when the robot uses deictic and/or iconic gestures.

The analysis was a doubly multivariate repeated measures ANOVA, with the translation scores (average of L2 to L1, and L1 to L2 translation tasks) and comprehension task scores as dependent variables, condition as independent variable, and age as covariate. The scores on the translation tasks were combined for all target words, which means that every participant had a score in the range of 0–34 (0.5 for each correctly translated word on one of the two translation tasks). The score on the comprehension task ranged from 0–54 (18 target words, 3 trials per word), where the chance of guessing correctly was 18 (33%), because every trial included the correct answer and two incorrect distractor items.

5.2.2. *Results*

Figure 10 shows the scores on the translation tasks of the immediate and delayed post-tests plotted against the participants' age in months at the start of the experiment. A linear fit to these data showed a steeper curve for the experimental condition where the robot used iconic gestures, that starts at a lower score on the translation tasks for younger children compared to the other experimental conditions, while it ends at a higher score than the other conditions for the older children in the study. This pattern did not emerge for the comprehension task, which is shown in Appendix C.

A doubly multivariate repeated measures ANOVA, with translation scores (combined into one score for both translation tasks) and scores for the comprehension task as dependent variables, condition as independent variable and children's age in months at the time of the pre-test as covariate, showed a significant effect of age for scores on the translation tasks, $F(1, 189) = 6.13, P = 0.01, \eta_p^2 = 0.03$, where older children in the study showed higher scores on the translation tasks of the post-tests than younger children. This effect was not significant for the comprehension task, $F(1, 189) = 1.24, P = 0.27, \eta_p^2 = 0.007$.

To further examine whether this effect held for all experimental conditions, we split the dataset and ran the aforementioned ANOVA per condition, with the translation scores and comprehension scores as dependent variables, and age as covariate. This showed the same significant effect of age for scores on the translation tasks, but only for the experimental condition where the robot used iconic gestures, $F(1, 52) = 4.59, P = 0.04, \eta_p^2 = 0.08$. No significant effects were found for the comprehension task, nor for any of the tasks in the other three conditions (all P-values in range $[0.26, 0.56]$).

The results of this analysis show that the older children in our study performed better on the translation (language production) tasks than the younger children, but only if the robot used iconic gestures while the children were learning the English words. Because this effect only shows in the experimental condition where the robot used iconic gestures, we postulate that older children may be better at understanding and making use of the robot's iconic gestures, compared to younger children. However, the effect of age should be interpreted with caution, because the effect size is relatively small.

5.3. Differences between semantic categories

Existing research suggests that iconic gestures for certain types of concepts (e.g. spatial concepts, motor events or items that are relatively concrete) contribute more strongly to learning than

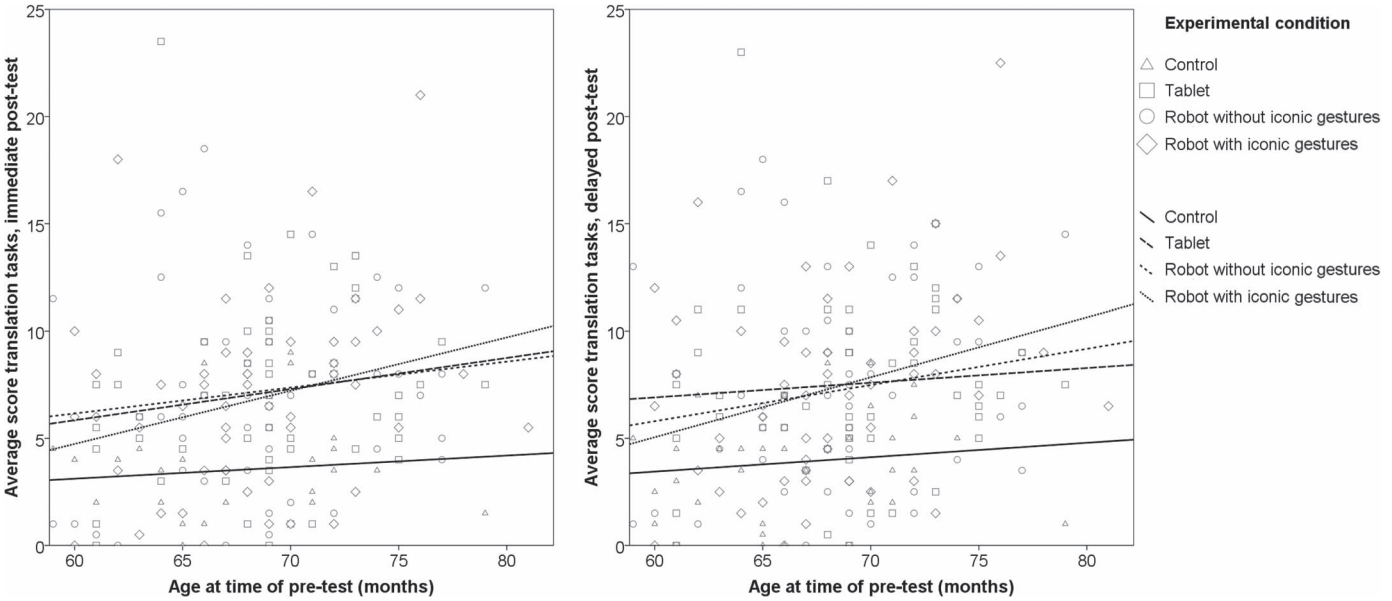

FIGURE 10. Linear fit to the post-test scores for the translation task per condition, by age.

gestures for concepts that are, for example, more abstract (de Nooijer *et al.*, 2013, Hostetter, 2011, Wakefield *et al.*, 2018). We therefore divided the English words into six semantic categories, and investigated whether there are any differences on average post-test scores between these categories, and if these can be attributed to the robot's use of gestures.

5.3.1. Analysis approach

For studying the differences between semantic categories, we included the tablet-only condition, where the robot was not physically present at all, and the two robot conditions (with and without iconic gestures), to see if the attention-guiding deictic gestures or the iconic gestures may have contributed to differences in children's learning outcomes for the different semantic categories. The 34 concepts for the translation tasks, and 18 concepts for the comprehension task, were divided into the same semantic categories used in the analysis of the gestures' comprehensibility (Subsection 5.1.1), and the post-test task scores were calculated for these semantic categories for the different experimental conditions. Scores on the translation tasks per child and per word were again either 0, 0.5, or 1, and for the comprehension task this was 0, 0.33, 0.66 or 1. These scores per child and per word were then averaged across children within the semantic categories, resulting in scores ranging from 0–1 for each category.

To check whether the differences between semantic categories were significant, we used a MANOVA with the scores on all six semantic categories, on the translation and comprehension tasks, as dependent variables (12 in total), and experimental condition (tablet-only, tablet + robot without iconic gestures, tablet + robot with iconic gestures) as independent variable. Furthermore, to test for an effect of age, in case differences occurred only for the older children in the sample, we ran the same MANOVA, including only the group of children who were at the average age of 5 years and 8 months or older (a *mean split*). This resulted in a subset of 38 children in the tablet-only condition, 32 in the tablet + robot without iconic gestures condition, and 29 in the tablet + robot with iconic gestures condition.

5.3.2. Results

Table 3 shows the average scores of all children on the post-test tasks, on the immediate and delayed post-tests, for the three experimental conditions. The table shows no large differences between conditions for any of the semantic categories. To test whether there were any statistically significant differences, we conducted a MANOVA with the post-test scores on the six semantic categories, on the translation tasks and the comprehension task, as dependent variables (12 in total), and experimental condition as independent variable. This showed no significant effect of experimental condition on children's performance on the semantic categories for the immediate post-test (all P-values in range $[0.11, 0.94]$), nor for the delayed post-test (all P-values in range $[0.23, 0.99]$).

Because we observed an effect of age, where the older children appeared to benefit more from the iconic gestures than the younger children in the study, we also present the average post-test task scores on the semantic categories of children that were at the average age of 5 years and 8 months or older (a *mean split*). These results are displayed in Table 4. This table shows differences between conditions, particularly on the translation tasks for the 'measurement words', where children

TABLE 3. Average translation and comprehension task scores (all 0–1) on semantic categories between conditions.

	Translation tasks			Comprehension task		
	Tablet	No iconic gestures	Iconic gestures	Tablet	No iconic gestures	Iconic gestures
Immediate post-test						
Counting	0.72	0.70	0.65	0.82	0.71	0.77
Comparatives	0.10	0.09	0.09	0.52	0.53	0.52
Operations	0.03	0.01	0.03	0.31	0.35	0.37
Measurement	0.10	0.09	0.12	0.56	0.57	0.58
Prepositions	0.03	0.04	0.03	0.42	0.41	0.36
Movement verbs	0.24	0.27	0.25	0.69	0.72	0.73
Delayed post-test						
Counting	0.78	0.71	0.71	0.69	0.67	0.69
Comparatives	0.12	0.12	0.09	0.52	0.56	0.53
Operations	0.01	0.00	0.02	0.69	0.62	0.67
Measurement	0.09	0.10	0.11	0.64	0.65	0.63
Prepositions	0.03	0.04	0.04	0.52	0.55	0.55
Movement verbs	0.26	0.26	0.26	0.46	0.43	0.46

in the condition with iconic gestures scored higher than the children in both other conditions.

To test whether there were significant differences between conditions, the same MANOVA was conducted for this subset of older participants, which showed a significant effect of condition for the 'measurement words' on the translation tasks on the immediate post-test, $F(2, 96) = 4.97, P = 0.009, \eta_p^2 = 0.09$, and for the translation tasks on the delayed post-test, $F(2, 96) = 5.85, P = 0.004, \eta_p^2 = 0.11$. For words related to 'operations', a significant effect of condition was found only for the translation tasks on the delayed post-test, $F(2, 96) = 3.60, P = 0.03, \eta_p^2 = 0.07$. No significant effects were found for categories other than 'measurement words' on the translation tasks of the immediate post-test (all P-values in range $[0.45, 0.96]$), and no significant effects were found for categories other than 'measurement words' and 'operations' on the translation tasks of the delayed post-test (all P-values in range $[0.11, 0.85]$). Furthermore, no significant effects were found for any of the semantic categories on the comprehension task, neither for the immediate post-test (all P-values in range $[0.11, 0.76]$) nor the delayed post-test (all P-values in range $[0.23, 0.99]$).

For the 'measurement words', a post-hoc analysis using Bonferroni correction shows a significant difference on the immediate post-test between the experimental condition with iconic gestures and the tablet-only condition ($M_{dif} = 0.96, P = 0.047$), and between the conditions with and without iconic gestures ($M_{dif} = 1.21, P = 0.01$). There was no significant difference between the tablet-only condition and the condition without iconic gestures ($M_{dif} = 0.25, P = 1.0$). For the delayed post-test, a post-hoc analysis using Bonferroni correction shows a significant difference between the condition with iconic gestures and the tablet-only condition ($M_{dif} = 1.19, P = 0.017$), and between the conditions with and without iconic gestures ($M_{dif} = 1.39, P = 0.006$), but not between the tablet-only condition and the condition without iconic gestures ($M_{dif} = 0.20, P = 1.0$).

The post-hoc tests for the 'operations words' on the delayed post-test showed no significant differences between the condition without iconic gestures and tablet-only condition ($M_{dif} = 0, P = 1.0$), between the condition with iconic gestures and the tablet-only condition ($M_{dif} = 0.17, P = 0.055$) or between the condition with iconic gestures and the condition without iconic gestures ($M_{dif} = 0.17, P = 0.07$). This is likely due to a floor effect, as shown by the .00 scores in the tablet-only condition and the condition without iconic gestures. Scores that are significantly different from the other experimental conditions have been marked in boldface in Table 4.

In summary, by comparing between experimental conditions we investigated whether the robot's physical presence, and its use of iconic gestures in particular, improved learning outcomes for specific semantic categories of words. When including all participants in the study, no differences between conditions were found for the semantic categories. However, after only including the older children in the study—those that appeared to be able to take advantage of the robot's gestures, as seen in Subsection 5.2—we observe that the robot's iconic gestures were mostly beneficial to learning the 'measurement words' (e.g. *big*), and they may have contributed to learning words pertaining to 'operations' (*add*, *take away*) as well.

5.4. Gesture reenactment

In several studies that report a positive contribution of iconic gestures to learning, participants were asked to not only observe, but to also perform the gestures themselves

TABLE 4. Average translation and comprehension task scores (all 0–1) on semantic categories between conditions, for children that were at least the average participant age of 5 years and 8 months (*mean split*). Values in boldface are significantly higher than in the other experimental conditions.

	Translation tasks			Comprehension task		
	Tablet	No iconic gestures	Iconic gestures	Tablet	No iconic gestures	Iconic gestures
Immediate post-test						
Counting	0.77	0.75	0.74	0.84	0.74	0.86
Comparatives	0.10	0.09	0.10	0.54	0.59	0.53
Operations	0.02	0.02	0.04	0.28	0.33	0.39
Measurement	0.08	0.06	**0.16**	0.55	0.55	0.61
Prepositions	0.04	0.04	0.03	0.39	0.39	0.35
Movement verbs	0.25	0.27	0.27	0.72	0.71	0.75
Delayed post-test						
Counting	0.80	0.77	0.81	0.67	0.65	0.67
Comparatives	0.12	0.14	0.09	0.52	0.56	0.55
Operations	0.00	0.00	0.04	0.70	0.61	0.68
Measurement	0.08	0.06	**0.18**	0.66	0.66	0.66
Prepositions	0.03	0.04	0.05	0.52	0.54	0.58
Movement verbs	0.27	0.26	0.29	0.46	0.43	0.49

(Cook *et al.*, 2008, de Nooijer *et al.*, 2013, Repetto *et al.*, 2017, Tellier, 2005, 2008). We assume that this could lead to a stronger grounding effect of the new vocabulary in existing sensorimotor experiences.

5.4.1. *Analysis approach*

To investigate whether children that spontaneously reenacted the gestures benefited more from them than children who did not perform gestures themselves, we annotated these reenactment events and compared them with the children's learning outcomes. This was done by reviewing the recordings of the interactions of all children that were in the experimental condition where the robot used iconic gestures ($N = 54$), and noting down every occurrence of reenactment including the timestamp within the video and the concept that was reenacted. For feasibility reasons, this annotation was only done for the first lesson, with the underlying assumption that this would help to identify the subgroup of 'reenacting children', and thereby give a representative idea of how much reenactment actually took place during the entirety of the experiment. Furthermore, in the last two lessons the children were prompted to enact a number of action verbs, which in the experimental condition with iconic gestures essentially means that children were actively requested to reenact the gestures, and therefore these lessons could not be included in an analysis of spontaneous reenactment. Due to technical issues, the robot did not gesture during the first lesson for one of the children, therefore we had to exclude this child from the analysis, resulting in 53 observed sessions. The relatively small sample size (for the number of words), low number of words learned overall, and the fact that reenactment was prompted in only two of the lessons means that we cannot perform analyses on the level of each word. The results should therefore be considered a first exploration of gesture reenactment and its effects on learning outcomes.

5.4.2. *Results*

In total, 37 out of the 53 children (70%) reenacted at least once during the first lesson. A total of 498 reenactment events were observed in this first lesson. When children reenacted, they did this 13 times on average ($SD = 13$), out of a minimum of 60 gestures performed by the robot, depending on the number of times the robot had to repeat a task. Figure 11 shows the frequency distribution of how often children reenacted the gestures and the frequency distribution of how many different concepts (out of 6) children reenacted during the first lesson. To see whether the act of imitating the iconic gestures from the robot relates to learning outcomes, we calculated the Pearson correlation between number of reenactments in lesson one and test scores on the comprehension and translation tasks. We focused on number of reenactments rather than a binary measurement (reenacted or not) to obtain a more precise estimate that reflected the variation in the sample more closely and, as such, may be more likely to show relationships with children's learning.

Table 5 shows the results of this correlation analysis. The correlation was not significant for the translation tasks nor for the comprehension task, in neither the immediate nor delayed post-test. In Appendix D we include a figure with each child's test scores on the vertical axis, and the number of times they reenacted during lesson one on the horizontal axis, showing no discernible pattern indicating a relationship between the number of reenactments during the first lesson, and children's learning outcomes. There was also no significant correlation between the children's age at the time of the pre-test, and the

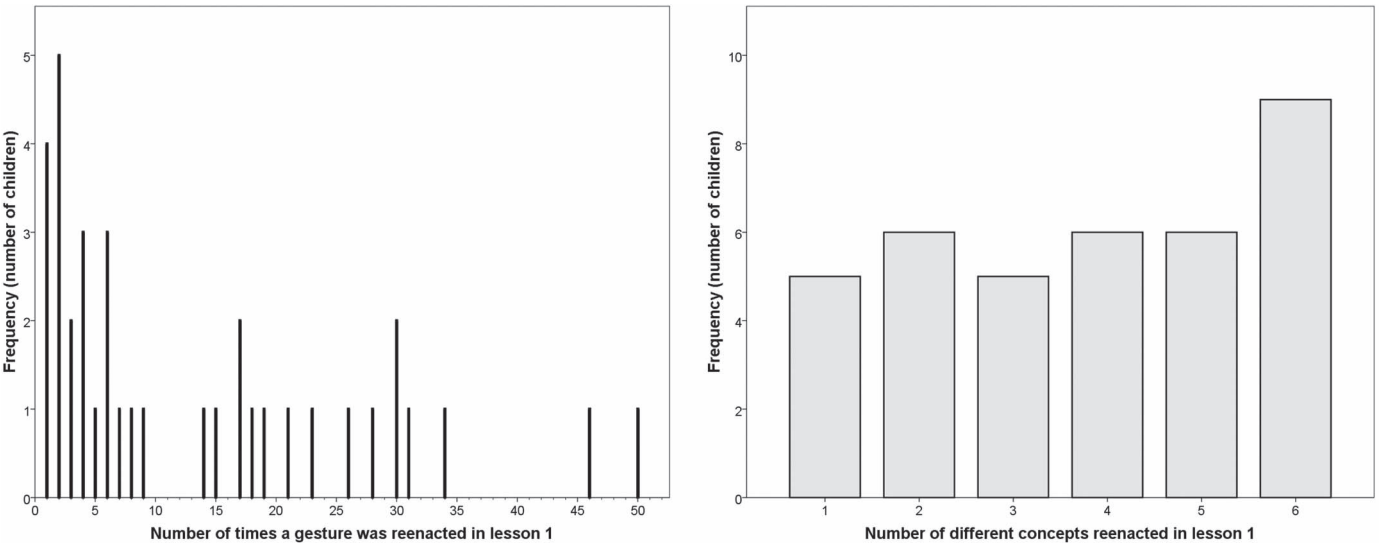

FIGURE 11. Left: Number of children (y-axis) that reenacted a certain number of times (x-axis) during the first lesson. Right: Number of children (y-axis) that reenacted a certain number of unique concepts (x-axis) during the first lesson. Only children that reenacted at least once are shown ($N = 37$; 16 did not reenact).

TABLE 5. Correlation between gesture reenactment and learning outcomes.

	r	*P*
Immediate post-test, translation tasks	-0.13	0.37
Immediate post-test, comprehension task	-0.09	0.53
Delayed post-test, translation tasks	-0.12	0.41
Delayed post-test, comprehension task	-0.17	0.24

number of reenactments during the first lesson, $r = -0.18, p = 0.19$.

Our investigation of spontaneous gesture reenactment shows that a relatively large number of children reenacted the robot's gestures during the first lesson (70%), compared to our previous experiences with running similar studies. However, reenactment did not appear to relate to learning outcomes, as there was no significant correlation between the number of reenactments in lesson one and the learning outcomes on the post-tests. In addition, the likelihood that a child in the study reenacted the robot's gestures did not appear to be linked to their age.

6. DISCUSSION

Existing literature in gesture studies and human-robot interaction suggests that iconic gestures, performed by humans or by robots, are able to support second language tutoring. However, our previous study (de Wit *et al.*, 2018) and the study that formed the basis of this paper (Vogt *et al.*, 2019) have shown mixed results, where in the case of our previous study the robot's iconic gestures did contribute to learning, while in the current study they did not. Therefore, in this paper we set out to explore a number of factors that may relate to the successful application of robot-performed gestures in second language tutoring. Concretely, we examined the importance of the design, and subsequent comprehensibility of the gestures (Q1), the age of the learners (Q2), differences between semantic categories of vocabulary words (Q3), and spontaneous gesture reenactment (Q4). In the following sections, we will address these subquestions, and infer guidelines for the design of robot-performed iconic gestures, focusing specifically on applications in (second language) education.

6.1. Design and comprehensibility of the robot's gestures

While reflecting upon the design of the robot's gestures, as well as their integration in the overall tutoring system, we identified several differences compared to our previous study (de Wit *et al.*, 2018). First, the English vocabulary words included in the current study were more complex, diverse, and abstract than the animal names that were used previously. These words may have been more difficult for children to learn, as seen in the small number of words learned in general, and the resulting gestures were less iconic than those from our previous study. A gesture for a concept such as *most* (shown in Figure 6), for example, will be more difficult to comprehend than a gesture that displays the act of riding a horse.

In addition, the positioning of the robot may have affected the clarity of these gestures. While in the previous study the robot was standing across from the child, in the current set-up it was sitting close to the child, at an angled position. This limited the robot to only use its upper body, and it changed

the perspective from which children were able to observe the gestures, which may have negatively affected their clarity. Concretely, we have seen that children misinterpreted gestures, as they were incorrectly mimicking them, for example by holding up their entire hand or showing three fingers for the word *two*. As a result of these factors, the gestures in the present research were likely more difficult to understand than those used in our previous study (de Wit *et al.*, 2018), in which the gestures had a higher degree of iconicity, the robot was positioned facing the child and the robot was able to use its full body to perform the gestures.

Although the gestures were designed based on recordings from an elicitation procedure, this procedure was conducted with adults rather than children from the same age group that would end up observing (and having to interpret) the gestures. Because children tend to perform gestures differently than adults do (Sekine *et al.*, 2018), it is conceivable that they also understand gestures that were produced by their peers better than those produced by people from a different age group. In future work we propose to take a more iterative approach to the design of gestures, including more frequent evaluations and revisions—with the target demographic, in this case, children—before integrating the final versions into the tutoring interaction.

The online evaluation with adults of the gestures shows that there are differences in the comprehensibility of the gestures, both between and within the semantic categories. As we observed while conducting the robot experiment with children, the gestures for counting words were often misinterpreted because of the NAO robot's inability to move its fingers independently. More generally, several of the iconic gestures that were originally recorded from human performers had to be reinterpreted to accommodate the robot's physical limitations and its positioning. We suspect that these reinterpreted versions were harder to understand than the original versions, but it would be interesting for future work to quantify, by means of empirical studies, this effect of redesigning gestures on their comprehensibility.

We did not observe a clear link between the gestures' comprehensibility and children's performance on learning the corresponding L2 words. It would be an interesting avenue for future research to study more closely this link between the quality, in terms of comprehensibility, of robot-performed gestures and how this relates to learning outcomes. We would then consider conducting the gesture evaluation study with children belonging to the same age group that would end up interacting with the robot. However, it might be difficult for younger children to correctly identify abstract words such as *big* and *take away* without any context other than the gesture, and they cannot be asked to judge the clarity and naturalness of the gestures. We therefore intend to explore alternative ways of conducting these rating studies with children in the future, perhaps in a game-like setting and using the child's L1 to provide context.

From this evaluation study we also found a moderate and significant correlation between the comprehensibility and clarity of the gestures, as well as their clarity and naturalness. Naturalness and comprehensibility also correlated moderately, but this was not significant ($p = .06$). Future research could look further into the nature of these relationships, to investigate how particular design aspects of gestures can be used to make the robot appear, for example, more human-like. It can be beneficial that a robot is perceived as human-like, as research has shown that this could, in turn, lead to the robot being perceived as warmer and more competent, which then leads to increased feelings of trust (Christoforakos *et al.*, 2021). Robots that look and behave in a human-like way are generally also seen as more likeable, and are more easily accepted by the people interacting with them (Roesler *et al.*, 2021).

The results of our anthropomorphism questionnaire, which are presented in more detail in van den Berghe *et al.* (2021a), indicated that in the current study children on average assigned more human-like traits than machine-like traits to the robot, although there were large individual differences. In addition, we did not see a difference in anthropomorphism between the experimental condition where the robot used deictic and iconic gestures, and the condition where it only used deictic gestures. This could mean that deictic gestures alone are enough to elicit higher degrees of anthropomorphism, but it is also possible that iconic gestures that are more natural or clear can further boost this effect. Our research further showed that the degree to which a robot tutor is seen as human-like by children correlates with the children's learning outcomes (van den Berghe *et al.*, 2021a), which leads us to believe that a robot that is perceived as human-like could be more successful as a (peer) tutor than one that is perceived as a toy or an artificial entity. It would therefore be interesting to explore which aspects of the robot's gestures lead to higher ratings on naturalness and clarity, as this could result in increased anthropomorphism, and perhaps, in turn, better learning outcomes.

Next to the design of the gestures themselves, and the limitations caused by the positioning of the robot, there are factors related to the integration of the iconic gestures into the ITS that could further explain why it may have been difficult for children to understand the gestures. For instance, the role that the tablet played within the overall interaction was smaller in our previous study compared to the current set-up. In de Wit *et al.*, (2018), the robot was the instructor during a game of 'I spy with my little eye' and children only had to select the correct image out of a number of answer options on the tablet. In the current experiment, children were asked to perform relatively complex tasks such as dragging objects in a three-dimensional virtual space. It is possible that children found these tasks more difficult, thereby drawing their attention away from the robot and its gestures. In addition, this could have increased cognitive load, resulting in less cognitive effort available to process the robot's gestures. An evaluation of the usability and user experience of the ITS also highlighted several issues that

may have negatively affected the quality of the interaction, some of which occurred more frequently or even exclusively in the experimental condition with gestures (de Wit *et al.*, 2019). Finally, the robot's gesturing interrupted the flow of the interaction. In order to time the motions so that the robot's pronunciation of the L2 words would coincide with the most salient part (the stroke) of the gesture, we introduced breaks in the robot's speech. Combined with the fact that 50 to 60 gestures were included in each lesson, this made the duration of the sessions substantially longer. As a result, the gestures had to maintain children's attention for a prolonged period of time. Research also indicates that a robot that gestures too frequently could be perceived as confusing and irritating (Pollmann *et al.*, 2020), although this was found with adults and it is as of yet unclear how different gesturing frequencies by robots are perceived by children. Additionally, the robot performed the same gesture for a particular concept every time, so it is possible that children got bored with seeing an identical motion ten times. Although the same limitations apply to our previous study (de Wit *et al.*, 2018), in the present study the interaction was more narrative-based, where the activities that the robot and child engaged in were linked to an overarching story line, compared to the more repetitive game of 'I spy with my little eye' used in the previous study. During this previous study, the gestures were also repeated less frequently and repetitions were spread out more over time.

6.2. Gestures and the effects of age

The fact that the older children in our study appeared to be able to understand and make use of the iconic gestures while the gestures seemed to have an adverse effect on younger children leads us to believe that either the gestures were too difficult or unclear for the younger participants in our study, or that younger children experienced some form of cognitive overload either due to the complexity of the interaction or the effort required to engage in learning second language vocabulary in combination with having to understand the gestures. Kennedy *et al.* (2015) also postulated that a robot's social behavior could add to cognitive load, making it more difficult for children to focus on the task. Cognitive overload may have distracted the children from the (phonetic elements of the) robot's speech as it was practicing the L2 words with them. It is worth noting that the effect size of age in the current study was relatively small, but so were the age differences (all children were approximately 5–6 years old). To further investigate this effect, and to explore which factors may have affected the results, we have recently conducted a follow-up study where we returned to the original, single session experiment from our previous work (de Wit *et al.*, 2018). We replaced the animal names with a more diverse set of concepts, and based the gestures on examples from a dataset of human-performed gestures—mostly performed by children and teenagers (de Wit *et al.*, 2020). Interestingly, in this study we observed a similar effect where older children (6 years old) did benefit from gestures, while younger children (4 years old) appeared to experience an adverse effect, causing them on average to learn fewer words than children their age in the experimental condition where the robot did not use gestures. The effect sizes in this case were larger, which could be attributed to the broader age range of 4–6 years old, the design of the study (e.g. single session compared to longitudinal), or the different measurement instruments (comprehension task, measured as pre-test and post-test, compared to translation tasks and a comprehension task measured as post-test).

From both the current study and the follow-up study it appears that there is a certain (cognitive) development that occurs between the age of 5 and 6, where children start being able to take advantage of the robot's gestures. Although literature indicates that we rely on gestures from a young age onward, it also shows that it takes time to fully understand and take advantage of them (Novack *et al.*, 2015, Stanfield *et al.*, 2014). Research by Stites & Özçalışkan (2017) further highlights that several aspects of gesture and speech change around the age of the participants in our study (5–6 years old). For example, they showed that children rely on gestures to support their speech when telling a narrative until the age of six, after which they start being able to use speech without support from gestures. It is therefore still possible that either the combination of foreign language learning and having to interpret gestures, or the multimodal interaction with a robot and a tablet may be too challenging for younger learners. This is further supported by a related study (van den Berghe *et al.*, 2021b), where we found that children with better selective attention (as measured using a visual search task; Mulder *et al.*, 2014) scored significantly higher on the post-tests if the robot used iconic gestures, compared to children with worse selective attention. It could also be the case that the children in our study differed in their ability to understand these two types of symbolic media—the robot's gestures and the depictions on the tablet screen (DeLoache, 2004). In future research we intend to run a gesture experiment with the robot but without a language learning component, in order to investigate whether this effect of age is indeed related to understanding the gestures, or to the cognitive effort needed to engage in the language learning interaction.

The effects of age on learning gain in the experimental condition with iconic gestures are only observed with results on the translation tasks, and not the comprehension task. This could either be because the gestures support children in acquiring a specific type of language skills (productive rather than receptive), or it could be due to the design of the tasks. Both our previous study (de Wit *et al.*, 2018) and the recent follow-up study (de Wit *et al.*, 2020) used only a (differently designed) comprehension task, and both found a positive effect of gestures on learning, either for all ages (de Wit *et al.*, 2018) or, similar to the translation task in the present study, with

age as a covariate (de Wit *et al.*, 2020). It is possible that the fact that only half of the concepts were included in the current comprehension task may have affected the quality of the measurements.

6.3. Differences between semantic categories

Existing literature indicates that gestures might be more effective at supporting learning of specific word types, such as spatial concepts or motor events (Hostetter, 2011), or verbs in general (Wakefield *et al.*, 2018). We therefore compared the percentage of correct answers on the post-tests for words belonging to the different semantic categories between experimental conditions (presented in Table 3). However, it appeared that children in the experimental condition with iconic gestures did not learn different types of vocabulary words than children that were in the other experimental conditions (tablet only, or robot without iconic gestures). For the counting words, it is conceivable that children already knew these words before participating in the study, which would explain why they score well on these words even though the gestures were not recognized by adult participants in the comprehensibility rating study. This is further supported by the fact that there were no differences between experimental conditions, and by children's performance on the pre-test translation task (L2 to L1 only), where they generally scored well on the counting words. This does not apply, however, to the movement verbs, of which the gestures received high comprehensibility scores, and for which children had relatively high post-test, but not pre-test scores. Children in the experimental condition with iconic gestures did not score better than those in the other conditions, therefore these words in general seem to have been relatively easy for children to learn compared to other semantic categories. This may be supported by the fact that the children in all experimental conditions were asked to act out these movements during the lessons.

Because only the older children in the study benefited from the robot's use of iconic gestures, we performed the same analysis on the subset of 99 out of 194 children that were older than the average age of the entire group of participants (5 years and 8 months); Table 4. For this group we do see a difference in performance on the translation tasks: Children who interacted with the robot that performed iconic gestures knew more words from the 'measurement' category (*big, small, heavy, light, high, low*), compared to children in the other experimental conditions. With the exception of *heavy*, these also had high comprehensibility scores, and above average ratings on clarity and naturalness in the gesture rating study (Appendix A). Because the group of older participants within the experimental condition with iconic gestures is relatively small ($N = 24$), further research is needed to verify whether indeed gestures are more useful for certain types of concepts than others.

6.4. Gesture reenactment

Research on the potential benefits of not only observing but also reenacting or mimicking gestures is scarce, but initial findings indicate that this can indeed lead to better learning outcomes compared to merely observing others produce the gestures (Cook *et al.*, 2008, de Nooijer *et al.*, 2013, Tellier, 2005, 2008). To our surprise, in the current analyses, 70% of the participants reenacted at least one gesture during the first lesson, without being prompted to do so. This is in stark contrast to our other studies with robot-performed gestures, where virtually no reenactment took place even though the participants were from the same age group as in the current study. Children might be more likely to imitate the robot's movements when it is positioned in a similar way to them: in this case both the child and the robot were sitting on the floor. The robot was also in relatively close physical proximity to the children, and it performed small gestures compared to the full-body movements in previous studies. Furthermore, there was a short pose imitation game included in the group introduction of the robot, which the children may have remembered during subsequent interactions. Another potential reason for the more frequent reenactment of gestures is the word repetition task, where the robot requested the child to verbally repeat one of the English terms. This task was not included in our previous studies, and while introducing this task the robot would also gesture, which may have inspired the child to accompany his or her verbal repetition with a gesture as well. Generally speaking, it is possible that children in the current study formed a stronger relationship with the robot, compared to the previous study. This could be due to a multitude of factors [see, e.g. van Straten *et al.*, 2020, for a review on child-robot relationship formation], such as the aforementioned physical proximity, and positioning the robot as a peer. Research suggests that familiarity with the demonstrator plays a role in whether children are likely to imitate behavior (e.g. Shimpi *et al.*, 2013).

Contrary to what is found in human gesture studies, in the current study we did not find a relationship between reenactment of the robot's gestures and learning outcomes. This may be due to the relatively small sample size (of reenacting children), or the relatively small effect of gestures in general. In the future we aim to study the role of gesture reenactment in social robot tutoring in a more structured way.

6.5. Strengths, limitations and future work

With this work we continue our line of research into robot-performed gestures and their effects on children's acquisition of second language vocabularies. We focused on the specific domain of second language tutoring, and within this domain on a particular set of English words for which gestures were developed. Based on this study alone, we cannot conclude that our findings will generalize to a broader range of (educational) domains, user types (e.g. adults), and robot platforms without

performing additional research. However, the results of the present study find support in existing research into human-robot interaction and gestures in general, which leads us to believe that this work is representative of the current state of social robots and robot-performed gestures. It is important to note that the original study, from which the data were used, was not designed with the aim to study the factors that are presented in this paper. Rather, these explorations were conducted post-hoc. Therefore, in the future we aim to conduct several follow-up studies to investigate these individual factors in more detail. It is also conceivable that there are factors other than the four currently covered, such as a child's prior exposure to iconic gestures (in an educational context) in everyday life, that have an effect on children's learning outcomes.

The focus of this study was on children's learning outcomes, but there may be additional effects such as engagement with the robot or with the educational content, and perception of the robot that have not yet been analyzed. We believe that these aspects of human-robot interactions are important to consider, and are planning to include these in future work. To further explore how the children were able to direct their attention during the interaction with the robot, and whether they were distracted by the presence of the tablet, we could examine their gaze patterns. Gaze has been shown to relate to engagement in child-robot interactions, and large individual differences are observed in this context (de Haas *et al.*, 2021). These individual differences in engagement and gaze behavior could explain differences in learning outcomes and the effectiveness of robot-performed gestures as well.

Furthermore, we wonder to what extent the inclusion of a tablet device has affected our results, especially since the content that was shown on the tablet was designed specifically for this study. It would be interesting to conduct a similar language learning study either with existing educational software or without the tablet device present at all. This might reduce cognitive load, which we believe may have hampered the gestures' effectiveness for the younger children in the study. By removing the tablet, we would be able to focus on the quality of the child-robot interaction, and the role that iconic gestures play in supporting this interaction.

It could be argued that a limitation of our study is that the words were not presented in a random order, within and between the lessons. We did this because we wanted to maintain a more structured flow through the different lessons, where each session had a clear theme, and there was a gradual increase in the difficulty level of the concepts (e.g. from relatively easy counting words to more complex spatial relations). Inspection of the results by checking how children did on words pertaining to particular lessons revealed no such order effects, however.

The design of the robot's iconic gestures was based on examples recorded from human participants in an elicitation study. This is an improvement over designing these gestures using a researcher's frame of mind. However, these recordings did not take into account the physical limitations, nor the seated and angled positioning of the robot. In addition, the recordings and the evaluations of the robot's gestures were both conducted with adult, non-expert participants, even though children would end up interacting with the robot. It is possible that children have different preferences when it comes to gesture strategies, which were now not included in the design. Instead of iteratively refining the gestures based on multiple evaluations, due to time constraints we only evaluated the gestures once after the study had already taken place. While this still allowed us to control for the quality of individual gestures as a potential confound, it did not improve the quality of the gestures before they were used in practice. We have observed in related literature that a validation of the gesture's design prior to using them in a study often does not take place at all. Therefore, to further improve the quality of the gestures, we propose to conduct more frequent evaluations, and to include participants who have similar demographic characteristics to the intended target audience.

A further potential limitation lies in the NAO robot that was used. This particular robot has certain physical limitations, that may be overcome in the future as robots are becoming more elaborate in their motion degrees of freedom. This could lead to more detailed, human-like gestures, which may have a stronger effect on learning outcomes, as well as the learner's perception and level of engagement. Regardless of the robot's physical limitations, we believe that the gestures that were developed in the current study are unique within the entire set of 34 target words, and that the majority of them correctly conveyed the meaning of the word they depicted. However, we only verified each gesture's comprehensibility in relation to five other concepts in our evaluation study. In the future we aim to more thoroughly explore how the robot's physical limitations affect the range of gestures that can be created for the robot, and the comprehensibility of these gestures.

With the primary goal of this line of research being to find technological solutions to support education, it is important to maintain a critical attitude toward the application of robots in this particular domain. There may be alternative options, such as tablet interactions or interactions with virtual agents, that may be equally (or more) effective compared to a robot. By studying (iconic) gestures, we are confident that we are focusing on one of the defining properties of social robots and their physical presence in the educational context. Although we believe our results to be promising in favor of the use of social robots, there are relatively high costs involved with deploying robots at schools on a large scale. More research is therefore needed to assess to what extent robot-performed gestures, when optimally designed and implemented, can improve a robot's tutoring efforts, and how social robots compare to alternative technologies. In the present study we may have linked the robot and tablet too strongly: In the tablet only condition, children were still introduced to the robot prior to the experiment, and the robot's voice was then routed through the tablet's speakers

during the experiment, which may have given children the impression that they were interacting with an embodied agent (cf. Moreno *et al.*, 2001).

7. CONCLUSION

We report on the design of an ITS, consisting of a tablet device and a social robot, with which children completed seven lessons of second language vocabulary training. This system was used to investigate whether a social robot can be used as a second language tutor for young children, particularly focusing on the design of the robot's iconic gestures. In the original, preregistered analyses of the results of the study with this ITS, we observed no benefits of the robot's use of iconic gestures to learning. This was in contrast with our results from previous studies. Therefore, in the current paper we provide an extended follow-up analysis of the empirical results collected in the study, focusing on four factors that, based on literature, may play a role in the successful application of robot-performed iconic gestures to support learning. These factors included (1) the quality of the gesture's design (and subsequent comprehensibility of these gestures), (2) the age of the learner and how that may affect their ability to make use of the gestures, (3) differences in effectiveness of gestures depending on the concept that is being described and (4) whether the learner reenacted or imitated the robot's gestures.

We found that, in the current study, gestures that were rated as more comprehensible by adults did not lead to better learning outcomes for children. The age of the participants did play a role in the experimental condition where the robot used iconic gestures: older children in this condition showed better learning outcomes compared younger children. The older children particularly benefited from gestures pertaining to measurement type words, such as *small*. Reenactment of the robot's gestures did not lead to increased learning outcomes in the current study.

This work contributes to the field of human-robot interaction by highlighting potential factors—gesture comprehensibility, age, types of concepts that are referred to, and gesture reenactment—that could play a role in the effectiveness of a robot's use of iconic gestures in an educational context. Based on our findings, we list several observations and suggestions regarding the process of designing a social robot's iconic gestures, and integrating them in a (tutoring) interaction:

- The positioning of the robot (and viewer's perspective) may relate to the comprehensibility and reenactment of the gestures;
- An iterative design process is desirable, including gesture elicitation and frequent evaluation with the intended user group (in our case children);
- Reinterpretation of human gestures to accommodate the robot's physical limitations may result in a loss of meaning or comprehension—frequent evaluation is especially useful in these cases;
- Particularly in education, aiming for human-like gestures (as opposed to more exaggerated gestures) may be useful, as human-likeness could offer benefits to the learning process and outcomes;
- Context matters: If the interaction is complex (e.g. including a tablet), frequent gesturing might cause cognitive overload. Managing the learner's attention is important;
- It may help to gesture less frequently (compared to human gesturing), and to introduce variation in gestures (i.e. using several different gestures for one word);
- It appears that the design and implementation of robot gestures can either discourage or encourage spontaneous reenactment.

In light of the present research and its promising outcomes, in future work we intend to conduct a study where we focus specifically on investigating these four factors in more detail.

ACKNOWLEDGMENTS

The authors would like to thank all the schools, parents and children that participated in our experiment; Chrissy Cook for being the voice of the tablet; as well as Laurette Gerts, Annabella Hermans, Esmee Kramer, Madée Kruijt, Marije Merckens, David Mogendorff, Sam Muntjewerf, Reinjet Oostdijk, Laura Pijpers, Chani Savelberg, Robin Sonders, Sirkka van Straalen, Sabine Verdult, Esmee Verheem, Pieter Wolfert, Hugo Zijlstra and Michelle Zomers for their invaluable help with running the experiment. The authors would also like to thank Peggy van Minkelen for evaluating the quality of the robot's gestures and for providing feedback on a draft version of the manuscript.

FUNDING

H2020 L2TOR project (grant 688014); Tilburg center for Cognition and Communication 'TiCC' at Tilburg University (the Netherlands).

REFERENCES

Ahmad, M. I., Mubin, O. and Orlando, J. (2016) Understanding behaviours and roles for social and adaptive robots in education: teacher's perspective. *In Proc. 4th International Conference on Human Agent Interaction*, 297–304. Association for Computing Machinery, New York, NY, United States.

Alemi, M., Meghdari, A., and Ghazisaedy, M. (2015) The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *Int. J. Soc. Robot.*, 7, 523–535.

Alibali, M. W. and Nathan, M. J. (2007) Teachers' gestures as a means of scaffolding students' understanding: evidence from an early algebra lesson. *Video Research in the Learning Sciences*, 349–365. Routledge.

Aly, A. and Tapus, A. (2013) A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction. In *Proc. 8th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 325–332. IEEE Press.

Anzalone, S. M., Boucenna, S., Ivaldi, S., and Chetouani, M. (2015) Evaluating the engagement with social robots. *Int. J. Soc. Robot.*, 7, 465–478.

Bainbridge, W. A., Hart, J. W., Kim, E. S., and Scassellati, B. (2011) The benefits of interactions with physically present robots over video-displayed agents. *Int. J. Soc. Robot.*, 3, 41–52.

Barsalou, L. W. (2008) Grounded cognition. *Annu. Rev. Psychol.*, 59, 617–645.

Bartneck, C. and Forlizzi, J. (2004) A design-centred framework for social human-robot interaction. In *RO-MAN 2004, 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No. 04TH8759)*, pp. 591–594. IEEE.

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

van den Berghe, R., Verhagen, J., Oudgenoeg-Paz, O., van der Ven, S., and Leseman, P. (2019) Social robots for language learning: a review. *Rev. Educ. Res.*, 89, 259–295.

van den Berghe, R., de Haas, M., Oudgenoeg-Paz, O., Krahmer, E., Verhagen, J., Vogt, P., Willemsen, B., de Wit, J., and Leseman, P. (2021a) A toy or a friend? Children's anthropomorphic beliefs about robots and how these relate to second-language word learning. *J. Comput. Assist. Learn.*, 37, 396–410.

van den Berghe, R., Oudgenoeg-Paz, O., Verhagen, J., Brouwer, S., de Haas, M., de Wit, J., Willemsen, B., Vogt, P., Krahmer, E. and Leseman, P. (2021b) Individual differences in children's (language) learning skills moderate effects of robot-assisted second language learning. *Front. Robot. AI*, 8, 259.

Blatchford, P. and Russell, A. (2020) *Rethinking Class Size: The Complex Story of Impact on Teaching and Learning*. UCL Press.

Breazeal, C. L. (2004) *Designing Sociable Robots*. The MIT Press.

Bremner, P. and Leonards, U. (2016) Iconic gestures for robot avatars, recognition and integration with speech. *Front. Psychol.*, 7, 183.

Bremner, P., Pipe, A. G., Melhuish, C., Fraser, M. and Subramanian, S. (2011) The effects of robot-performed co-verbal gesture on listener behaviour. In *The 2011 11th IEEE-RAS International Conference on Humanoid Robots*, pp. 458–465. IEEE.

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

Chiat, S. (2015) Nonword repetition. In *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*, pp. 125–150. Multilingual Matters/Channel View Publications Ltd.

Christoforakos, L., Gallucci, A., Surmava-Große, T., Ullrich, D. and Diefenbach, S. (2021) Can robots earn our trust the same way humans do? a systematic exploration of competence, warmth and anthropomorphism as determinants of trust development in hri. *Front. Robot. AI*, 8, 79.

Cook, S. W., Mitchell, Z., and Goldin-Meadow, S. (2008) Gesturing makes learning last. *Cognition*, 106, 1047–1058.

Craenen, B., Deshmukh, A., Foster, M. E. and Vinciarelli, A. (2018) Shaping gestures to shape personalities: The relationship between gesture parameters, attributed personality traits and godspeed scores. In *The 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 699–704. IEEE.

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends Cogn. Sci.*, 8, 66–70.

van Dijk, E. T., Torta, E., and Cuijpers, R. H. (2013) Effects of eye contact and iconic gestures on message retention in human-robot interaction. *Int. J. Soc. Robot.*, 5, 491–501.

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robot. Auton. Syst.*, 42, 177–190.

Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003) A survey of socially interactive robots. *Robot. Auton. Syst.*, 42, 143–166.

Fridin, M. (2014) Kindergarten social assistive robot: first meeting and ethical issues. *Comput. Hum. Behav.*, 30, 262–272.

Gielniak, M. J. and Thomaz, A. L. (2012) Enhancing interaction through exaggerated motion synthesis. In *Proc. 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 375–382. ACM.

Glenberg, A. M. and Gallese, V. (2012) Action-based language: a theory of language acquisition, comprehension, and production. *Cortex*, 48, 905–922.

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M. and Breazeal, C. (2016) Affective personalization of a social robot tutor for children's second language skills. In *The Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, Palo Alto, California USA.

de Haas, M., Vogt, P., and Krahmer, E. (2021) When preschoolers interact with an educational robot, does robot feedback influence engagement? *Multimodal Technol. Interact.*, 5, 77.

Hald, L. A., de Nooijer, J., van Gog, T., and Bekkering, H. (2016) Optimizing word learning via links to perceptual and motoric experience. *Educ. Psychol. Rev.*, 28, 495–522.

Han, J.-H., Jo, M.-H., Jones, V., and Jo, J.-H. (2008) Comparative study on the educational use of home robots for children. *J. Inform. Process. Syst.*, 4, 159–168.

Henkemans, O. A. B., Bierman, B. P., Janssen, J., Neerincx, M. A., Looije, R., van der Bosch, H., and van der Giessen, J. A. (2013) Using a robot to personalise health education for children with diabetes type 1: a pilot study. *Patient Educ. Couns.*, 92, 174–181.

Hostetter, A. B. (2011) When do gestures communicate? A meta-analysis. *Psychol. Bull.*, 137, 297, 315.

Hostetter, A. B. and Alibali, M. W. (2008) Visible embodiment: gestures as simulated action. *Psychon. Bull. Rev.*, 15, 495–514.

Howley, I., Kanda, T., Hayashi, K., and Rosé, C. (2014) Effects of social presence and social role on help-seeking and learning. In *Proc. of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 415–422. ACM.

Huang, C.-M. and Mutlu, B. (2013) Modeling and evaluating narrative gestures for humanlike robots. In *Robotics: Science and Systems*, pp. 57–64.

Kanda, T., Shimada, M. and Koizumi, S. (2012) Children learning with a social robot. In *The 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 351–358. IEEE.

Kanero, J., Demir-Lira, E., Koskulu, S., Oranç, C., Franko, I., Küntay, A. C., and Göksun, T. (2018a) *How do robot gestures help second language learning? In Earli SIG 5 Abstract Book*.

Kanero, J., Geçkin, V., Oranç, C., Mamus, E., Küntay, A. C., and Göksun, T. (2018b) Social robots for early language learning: current evidence and future directions. *Child Dev. Perspect.*, 12, 146–151.

Kelly, S. D., Manning, S. M., and Rodak, S. (2008) Gesture gives a hand to language and learning: perspectives from cognitive neuroscience, developmental psychology and education. *Lang. Linguist. Compass*, 2, 569–588.

Kelly, S. D., McDevitt, T., and Esch, M. (2009) Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Lang. Cogn. Process.*, 24, 313–334.

Kendon, A. (2004) *Gesture: Visible Action as Utterance*. Cambridge University Press.

Kennedy, J., Baxter, P. and Belpaeme, T. (2015) The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. In *The 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 67–74. IEEE.

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E. and Belpaeme, T. (2017) Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proc. of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (), pp. 82–90. ACM.

Klahr, D., Triona, L. M., and Williams, C. (2007) Hands on what? The relative effectiveness of physical versus virtual materials in an engineering design project by middle school children. *J. Res. Sci. Teach.*, 44, 183–203.

Kory-Westlund, J. K., Dickens, L., Jeong, S., Harris, P., DeSteno, D. and 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*. Windesheim Flevoland, Almere, the Netherlands.

Lee, S., Noh, H., Lee, J., Lee, K., Lee, G. G., Sagong, S., and Kim, M. (2011) On the effectiveness of robot-assisted language learning. *ReCALL*, 23 25–58.

Li, J. (2015) The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *Int. J. Hum.-Comp. Stud.*, 77, 23–37.

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

McNeil, N. M., Alibali, M. W., and Evans, J. L. (2000) The role of gesture in children's comprehension of spoken language: now they need it, now they don't. *J. Nonverbal Behav.*, 24, 131–150.

McNeill, D. (1992) *Hand and mind: what gestures reveal about thought*. University of Chicago Press.

Moreno, R., Mayer, R. E., Spires, H. A., and Lester, J. C. (2001) The case for social agency in computer-based teaching: do students learn more deeply when they interact with animated pedagogical agents? *Cogn. Instr.*, 19, 177–213.

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

Mubin, O., Stevens, C. J., Shahid, S., Al Mahmud, A., and Dong, J.-J. (2013) A review of the applicability of robots in education. *J. Technol. Educ. Learn.*, 1, 13.

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

de Nooijer, J. A., Van Gog, T., Paas, F., and Zwaan, R. A. (2013) Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. *Acta Psychol. (Amst)*, 144, 173–179.

Novack, M. A., Goldin-Meadow, S. and Woodward, A. L. (2015) Learning from gesture: how early does it happen? *Cognition*, 142, 138–147.

Özgür, A., Johal, W., Mondada, F. and Dillenbourg, P. (2017) Windfield: learning wind meteorology with handheld haptic robots. In *Proc. of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (), pp. 156–165, ACM.

Pollmann, K., Ruff, C., Vetter, K. and Zimmermann, G. (2020) Robot vs. voice assistant: is playing with Pepper more fun than playing with Alexa? In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI '20*, pp. 395–397. Association for Computing Machinery, New York, NY, USA.

Pot, E., Monceaux, J., Gelin, R. and Maisonnier, B. (2009) Choregraphe: A graphical tool for humanoid robot programming. In *RO-MAN 2009: The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 46–51. IEEE.

Ramachandran, A., Huang, C.-M., Gartland, E. and Scassellati, B. (2018) Thinking aloud with a tutoring robot to enhance learning. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 59–68, ACM.

Repetto, C., Pedroli, E. and Macedonia, M. (2017) Enrichment effects of gestures and pictures on abstract words in a second language. *Front. Psychol.*, 8, 2136.

Robert, L. P., Alahmad, R., Esterwood, C., Kim, S., You, S., and Zhang, Q. (2020) A review of personality in human robot interactions. *Foundations and Trends® in Information Systems, vol. 4*, pp. 107–212, Now Publishers, Inc.

Roesler, E., Manzey, D., and Onnasch, L. (2021) A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Sci. Robot.*, 6, eabj5425.

Rohlfing, K. J. (2019) Learning language from the use of gestures. In Horst, J., von Koss Torkildsen, J. (eds), *International Handbook of Language Acquisition*, pp. 213–233. Routledge/Taylor & Francis Group.

Roth, W.-M. (2001). Gestures: their role in teaching and learning. *Rev. Educ. Res.*, 71, 365–392.

Rowe, M. L., Silverman, R. D., and Mullan, B. E. (2013) The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemp. Educ. Psychol.*, 38, 109–117.

Saerbeck, M., Schut, T., Bartneck, C. and Janse, M. D. (2010) Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1613–1622. ACM.

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., and Joublin, F. (2013) To err is human (−like): effects of robot gesture on perceived anthropomorphism and likability. *Int. J. Soc. Robot.*, 5, 313–323.

Sauppé, A. and Mutlu, B. (2014) Robot deictics: how gesture and context shape referential communication. In *Proc. of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 342–349, ACM.

Scassellati, B. (2002) Theory of mind for a humanoid robot. *Auton. Robots*, 12, 13–24.

Schlichting, L. (2005) *Peabody picture vocabulary test-III-NL*. Hartcourt Assessment BV, Amsterdam, the Netherlands.

Sekine, K., Wood, C., and Kita, S. (2018) Gestural depiction of motion events in narrative increases symbolic distance with age. *Language, Interaction and Acquisition*, 9, 40–68.

Shimpi, P. M., Akhtar, N., and Moore, C. (2013) Toddlers' imitative learning in interactive and observational contexts: the role of age and familiarity of the model. *J. Exp. Child Psychol.*, 116, 309–323.

Singer, I. and Gerrits, E. (2015) The effect of playing with tablet games compared with real objects on word learning by toddlers. *In Conference Proceedings ICT for Language Learning*, 255–259. Libreria Universitaria, Padova, Italy.

Stanfield, C., Williamson, R., and Özçalışkan, Ş. (2014) How early do children understand gesture–speech combinations with iconic gestures? *J. Child Lang.*, 41, 462–471.

Stites, L. J. and Özçalışkan, Ş. (2017) Who did what to whom? Children track story referents first in gesture. *J. Psycholinguist. Res.*, 46, 1019–1032.

van Straten, C. L., Peter, J., and Kühne, R. (2020). Child–robot relationship formation: a narrative review of empirical research. *Int. J. Soc. Robot.*, 12, 325–344.

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

Tellier, M. (2005) *How do teacher's gestures help young children in second language acquisition?* International Society of Gesture Studies, ISGS.

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

Toh, L. P. E., Causo, A., Tzuo, P.-W., Chen, I.-M., and Yeo, S. H. (2016) A review on the use of robots in education and young children. *J. Educ. Technol. Soc.*, 19, 148–163.

Valenzeno, L., Alibali, M. W., and Klatzky, R. (2003) Teachers' gestures facilitate students' learning: a lesson in symmetry. *Contemp. Educ. Psychol.*, 28, 187–204.

Vlaar, R., Verhagen, J., Oudgenoeg-Paz, O. and Leseman, P. (2017) Comparing L2 word learning through a tablet or real objects: what benefits learning most? In *Proceedings of the R4L workshop, at the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Association for Computing Machinery, New York, NY, USA.

Vogt, P., De Haas, M., De Jong, C., Baxter, P. and Krahmer, E. (2017) Child-robot interactions for second language tutoring to preschool children. *Front. Hum. Neurosci.*, 11, 73.

Vogt, P.*et al.* (2019) Second language tutoring using social robots: a large-scale study. In *The 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 497–505. IEEE.

Wakefield, E. M., Hall, C., James, K. H., and Goldin-Meadow, S. (2018) Gesture for generalization: gesture facilitates flexible learning of words for actions on objects. *Dev. Sci.*, 21, e12656.

de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., Krahmer, E. and 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. ACM.

de Wit, J., Pijpers, L., van den Berghe, R., Krahmer, E. and Vogt, P. (2019) Why ux research matters for hri: the case of tablets as mediators. In *Workshop on the Challenges of Working on Social Robots that Collaborate with People, at the ACM CHI Conference on Human Factors in Computing Systems (CHI2019)*. ACM.

de Wit, J., Brandse, A., Krahmer, E. and 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, HRI '20*, pp. 359–367. Association for Computing Machinery, New York, NY, USA.

Zaga, C., Lohse, M., Truong, K. P. and Evers, V. (2015) The effect of a robot's social character on children's task engagement: Peer versus tutor. In *International Conference on Social Robotics*, pp. 704–713. Springer.

A Detailed Results of the Gesture Rating Study with Adults

TABLE A6. Comprehensibility, clarity and naturalness ratings for each gesture (SD in parentheses). Chance level for comprehensibility is 0.17.

Concept	Semantic category	Comprehensibility 0–1	Clarity 1–5	Naturalness 1–5
One	Counting	0.35	2.35 (1.00)	2.41 (0.94)
Two	Counting	0.53	3.41 (1.33)	3.41 (1.00)
Three	Counting	0.24	4.12 (0.86)	3.88 (0.98)
Four	Counting	0.29	3.06 (1.20)	2.65 (0.86)
Five	Counting	0.35	2.06 (1.03)	2.35 (1.11)
More	Comparatives	0.59	3.24 (0.90)	3.12 (0.86)
Most	Comparatives	0.18	3.12 (1.05)	3.24 (0.90)
Fewer	Comparatives	0.65	3.12 (0.99)	3.18 (0.95)
Fewest	Comparatives	0.65	3.41 (1.12)	3.24 (1.03)
Add	Operations	0.94	2.53 (1.37)	2.53 (1.28)
Take away	Operations	1.00	3.53 (1.01)	3.35 (0.93)
Big	Measurement	0.94	3.94 (0.83)	3.65 (1.17)
Small	Measurement	0.76	3.82 (1.13)	3.71 (0.99)
Heavy	Measurement	0.29	3.88 (0.93)	3.65 (0.79)
Light	Measurement	0.88	3.47 (0.72)	3.41 (0.80)
High	Measurement	0.88	4.24 (0.83)	3.88 (0.78)
Low	Measurement	0.71	4.18 (0.73)	3.76 (0.66)
On	Prepositions	0.47	3.71 (0.92)	3.06 (0.90)
Above	Prepositions	0.76	3.65 (0.93)	3.18 (0.88)
Below	Prepositions	0.94	4.35 (1.06)	4.29 (0.69)
Next to	Prepositions	0.76	3.12 (1.36)	3.06 (1.03)
In front of	Prepositions	0.88	4.06 (1.03)	3.71 (0.92)
Behind	Prepositions	0.76	3.35 (1.22)	3.00 (1.00)
Left	Prepositions	0.82	4.29 (0.77)	4.06 (0.56)
Right	Prepositions	0.94	4.59 (0.62)	4.12 (0.78)
Falling	Movement verbs	0.88	4.53 (0.72)	4.29 (0.77)
Walking	Movement verbs	1.00	4.88 (0.33)	4.35 (0.93)
Running	Movement verbs	1.00	4.47 (1.01)	4.41 (0.62)
Jumping	Movement verbs	0.76	3.12 (1.32)	3.06 (1.09)
Flying	Movement verbs	0.94	4.41 (0.80)	3.65 (1.41)
Catching	Movement verbs	0.76	3.41 (1.00)	3.18 (1.01)
Throwing	Movement verbs	0.88	4.65 (0.61)	4.24 (0.97)
Sliding	Movement verbs	0.71	2.71 (0.92)	2.82 (0.81)
Climbing	Movement verbs	1.00	4.82 (0.39)	4.47 (0.51)

B Gesture Rating Study: Differences between Semantic Categories

TABLE B2. Paired samples t-tests to test differences between semantic categories. * indicates significant difference.

	M_{dif}	SD_{dif}	$t(16)$	p
Comprehensibility (0–1)				
Counting–comparatives*	-0.16	0.31	-2.14	.048
Counting–operations*	-0.62	0.24	-10.59	¡.001
Counting–measurement*	-0.39	0.26	-6.23	¡.001
Counting–prepositions*	-0.44	0.30	-6.11	¡.001
Counting–movement*	-0.53	0.27	-8.14	¡.001
Comparatives–operations*	-0.46	0.24	-7.90	¡.001
Comparatives–measurement*	-0.23	0.20	-4.83	¡.001
Comparatives–prepositions*	-0.28	0.24	-4.72	¡.001
Comparatives–movement*	-0.37	0.25	-5.98	¡.001
Operations–measurement*	0.23	0.16	5.99	¡.001
Operations–prepositions*	0.18	0.23	3.23	.005
Operations–movement*	0.09	0.15	2.50	.02
Measurement–prepositions	-0.05	0.19	-1.05	.30
Measurement–movement*	-0.13	0.16	-3.49	.003
Prepositions–movement*	-0.09	0.16	-2.25	.04
Clarity (1–5)				
Counting–comparatives	-0.22	0.66	-1.38	.19
Counting–operations	-0.03	0.78	-0.16	.88
Counting–measurement*	-0.92	0.56	-6.83	¡.001
Counting–prepositions*	-0.89	0.44	-8.35	¡.001
Counting–movement*	-1.11	0.44	-10.39	¡.001
Comparatives–operations	0.19	0.89	0.89	.39
Comparatives–measurement*	-0.70	0.48	-6.06	¡.001
Comparatives–prepositions*	-0.67	0.57	-4.82	¡.001
Comparatives–movement*	-0.89	0.57	-6.43	¡.001
Operations–measurement*	-0.89	0.88	-4.20	.001
Operations–prepositions*	-0.86	0.80	-4.44	¡.001
Operations–movement*	-1.08	0.71	-6.30	¡.001
Measurement–prepositions	0.03	0.50	0.26	.80
Measurement–movement	-0.19	0.40	-1.95	.07
Prepositions–movement*	-0.22	0.35	-2.63	.02
Naturalness (1–5)				
Counting–comparatives	-0.25	0.65	-1.57	.14
Counting–operations	0.00	0.74	0.00	1.00
Counting–measurement*	-0.74	0.61	-4.94	¡.001
Counting–prepositions*	-0.62	0.44	-5.73	¡.001
Counting–movement*	-0.89	0.58	-6.36	¡.001
Comparatives–operations	0.25	0.99	1.04	.31
Comparatives–measurement*	-0.49	0.52	-3.86	.001
Comparatives–prepositions*	-0.37	0.49	-3.10	.007
Comparatives–movement*	-0.64	0.68	-3.89	.001
Operations–measurement*	-0.74	0.89	-3.40	.004
Operations–prepositions*	-0.62	0.88	-2.90	.01
Operations–movement*	-0.89	0.87	-4.22	.001
Measurement–prepositions	0.12	0.30	1.63	.12
Measurement–movement	-0.15	0.41	-1.54	.14
Prepositions–movement*	-0.27	0.42	-2.67	.02

C Relation between age and post-test scores on the comprehension task

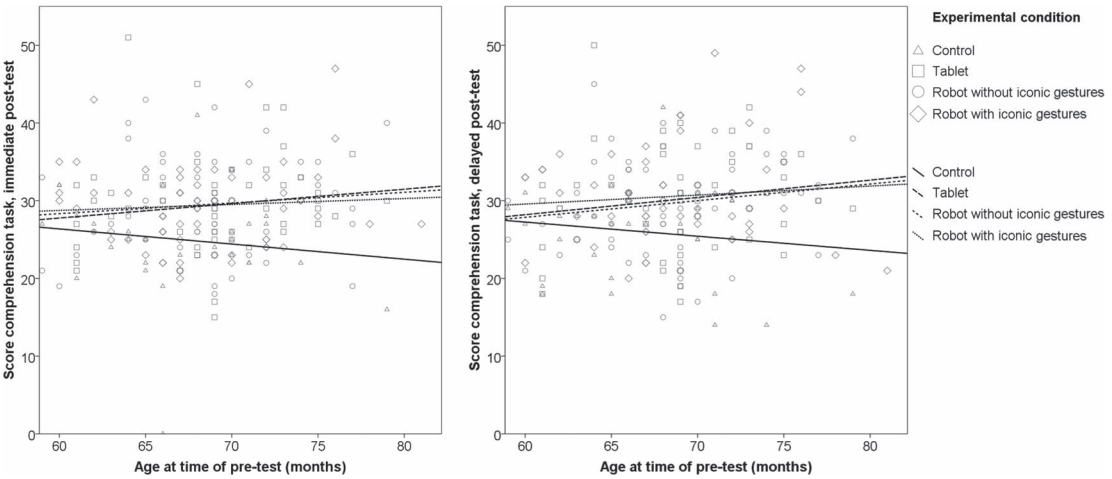

FIGURE C2. Linear fit to the post-test scores for the comprehension task per condition, by age (chance level is 18).

D Relation between reenactment and learning gain

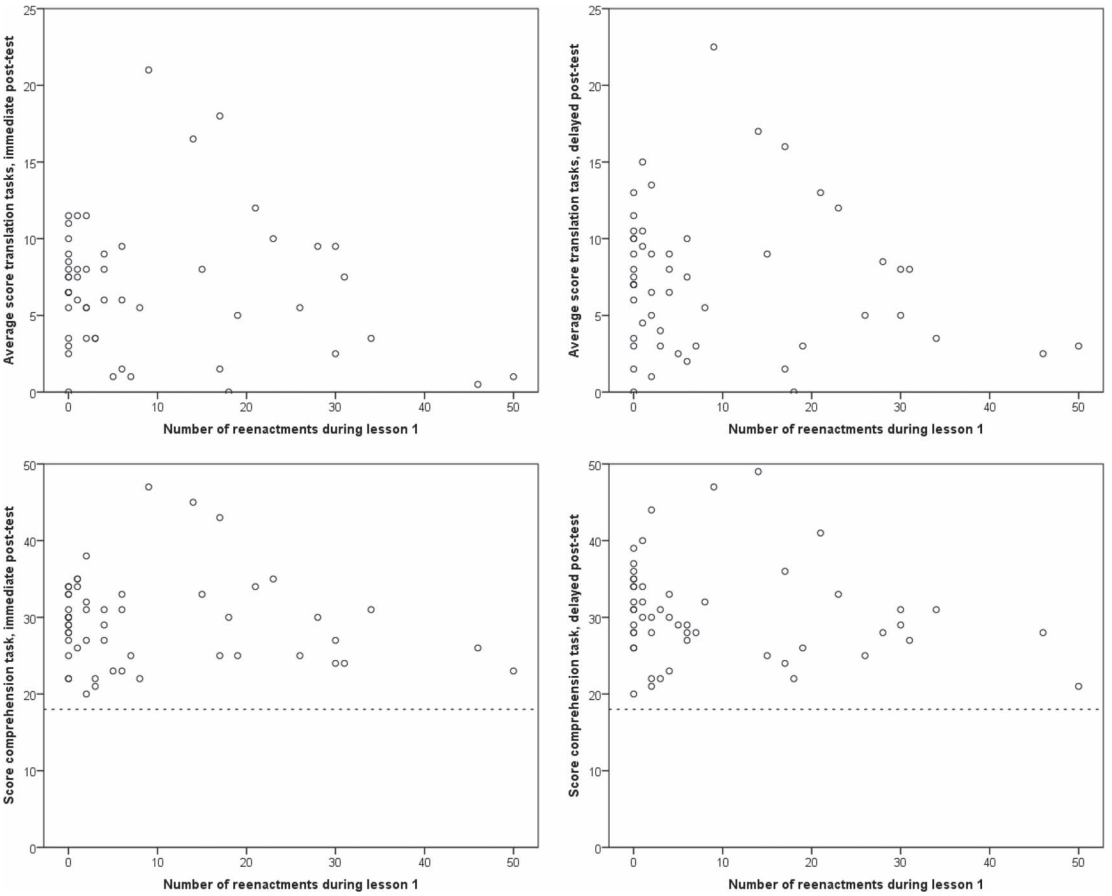

FIGURE D3. Post-test scores on the translation tasks (top) and comprehension task (bottom) plotted against the number of reenacted gestures in lesson 1. Chance level is 18 for the comprehension task.