Robotic Gesture Use in L2 Learning

The effects of repetition and variation in robotic gesture production on children's second language acquisition, engagement, and perceived anthropomorphism

Arold Brandse, ANR: 2017429 – u172532

Master's Thesis

Communication & Information Sciences — New Media Design

School of Humanities and Digital Sciences

Tilburg University, Tilburg

Supervisors: J.M.S de Wit, MSc & prof. dr. E.J. Krahmer

Second Reader: drs. J.M. van der Loo

August 2019

Acknowledgements

This thesis would never have come in to being without help and assistance of several people. I would first like to thank my supervisor, Jan de Wit. Without his help, nothing in this experiment would have come to pass. His enthusiasm for robots has made me grow new interests. I would also like to thank Emiel Krahmer for providing additional support and reading of a concept version of this thesis even during a vacation.

I would like to extend my gratitude to Jan de Wit, Reinjet Oostdijk and Martijn Faes for helping out with data collection in the overlap period between the two schools.

In addition, I would like to thank my family, friends and fellow students for their support. A special gratitude to Henrike Colijn for her great support and help with last-minute quality assurances.

I would also like to thank Cindy van Hemert from *Nutsbasisschool Teteringen*, and Marjolein Backx from *Basisschool De Stappen*, as well as both schools for their help and hospitality. Without these schools, this study would not have been possible.

Finally, I would like to thank all children and parents for their participation in this study, as well as their interest and enthusiasm.

Abstract

Many studies have looked specifically at the language tutoring capabilities of robots, yet few have looked at using the robot to portray gestures, even though gestures have been found to improve language learning performance. Of what little research there is, the findings show mixed results. The present study is based on one of these studies, specifically by de Wit et al. (2018), and focuses on the effects of gestures as well as variations in gesture production by robots on second language word learning, engagement and perceived anthropomorphism in young children. A three-group field experiment was set up where a robot employed either no gestures, a single repeated gesture for each unique target word, or a new gesture each time a word was presented. In total, 94 children (mean age of five years and three months) participated in this study. Based on a pre-registered analysis an overall learning result was found, however no differences were found between the three conditions. Engagement did see differences between conditions, with gesture conditions resulting into higher child-robot engagement than the non-gesture condition. A more thorough exploratory analysis revealed that age of the children played a large role in the learning results. Children aged five and six learned significantly more than children aged four. Trends were visible that they learned more from gestures as well, though these were not significant. In engagement too, differences between the two age groups were found. Interestingly, no differences were found in perceived anthropomorphism across the board. Implications of these findings are discussed at the end of this thesis.

Keywords: Robotics; Gestures; Language learning; Engagement; Anthropomorphism

Table of Contents

1. Introduction

In the last decade, the use of digital media tools classrooms has become increasingly more prevalent. In the Netherlands, many elementary schools have widely adopted digital learning environments, tablets, digiboards, serious games and many others. These tools provide numerous advantages such as data-collection on a child's long-term performance, communication between teacher and pupil, practice, instruction and variation in learning materials (Haelermans, 2017; van Elk, 2018). More recently, with advances in both research and technology, robots have started entering the world of digital education. Specifically, these types of robots are named *social robots*, and are defined as "A physical entity embodied in a complex, dynamic, and social environment sufficiently empowered to behave in a manner conducive to its own goals and those of its community" (Duffy, Rooney, O'Hare, & O'Donoghue, 2000, p. 4) They differ from static, industrial robots many people are familiar with and instead are much more dynamic, similar to humans. In essence, they are typically capable of some form of communication in such a way that it allows humans to form bonds with them. With this recent socialisation of robots, an entirely new area of study has spawned: Human-Robot Interaction (HRI), and while it is still a relatively young field of study, it has steadily been growing in recent years.

Particularly in the educational setting, research has been expanding on how social robots can help with the development of children, spawning a subfield of HRI: Child-Robot Interaction (cHRI). To study and optimise the capabilities of social robots, this field has been looking into how they can serve as intermediary learning tools. Amongst others, they have been found to assist with language learning and cognitive development (Mubin, Stevens, Shahid, Mahmud, & Dong, 2013; Slangen, Van Keulen, & Gravemeijer, 2011), aid in the development of problem-solving skills (Barak & Zadok, 2009), promote collaboration (Shimada, Kanda, & Koizumi, 2012; Varney, Janoudi, Aslam, & Graham, 2012), promote interest and motivation in (technological) subjects (Ruiz-del-Solar & Avilés, 2004), and improve learning in a varying range of subjects such as programming (Kazakoff, Sullivan, & Bers, 2013), evolution theory (Whittier & Robinson, 2007), and physics (Williams, Ma, Prejean, Ford, & Lai, 2007).

In recent years, an especial interest has been taken in robot-assisted language learning (RALL). As a key advantage over more traditional media such as computers, the physicality of robots allows for the handling of objects, as well as bodily movements and gestures that can aid in language learning (Mavilidi, Okely, Chandler, Cliff, & Paas, 2015; Rowe & Goldin-Meadow, 2009). This is interesting especially for language learning, as the use of gestures has strongly been linked to an increase in language learning (e.g., Macedonia, Müller, & Friederici, 2011; Macedonia & Von Kriegstein, 2012; Sueyoshi & Hardison, 2005; Tellier, 2008). Yet few robot studies have looked at the advantages of gesture use in RALL.

Of what little research there is, mixed results were found. A longitudinal study (Vogt et al., 2019) found no benefits in second language learning (referred to as L2, with L1 being one's native language) when robots performed gestures, while a single session study (de Wit et al., 2018) did find results. Vogt et al. (2019) alluded to what may have happened though: some types of gestures seemed to work better than others. A study by de Wit et al. (2019) showed indicators that may support this hypothesis, when attempting to use robots to capture and recognise gestures. Participants were asked to act out meanings of different words using gestures and body language in front of a robot. Amongst others, the results showed that different people have diverse ways of portraying these words (Figure 1). By extension, it may very well be possible that gestures can have several types of interpretations. If true, this could partially indicate why Vogt et al. (2019) found differences between gesture types in children: their young lifetime does not permit them to build up as large of a recollection of different types of gestures portraying certain words.

Figure 1. The generated dataset of the study showed various interpretations of the word 'guitar'. The first and second images show gestures of children, while the last two images show gestures by adults. Reprinted from de Wit et al. (2019) with permission.

To that effect, the current study's aim is twofold: First of all, it intends to add to the body of knowledge surrounding robotic gesture use in L2 learning. This study will look at how gestures impact children's L2 learning performance, how well they stay engaged with the robot, and finally their perception of how human-like the robot is (also referred to as anthropomorphism). Second, this study intends to further explore the different personal connotations humans have with certain words and gestures, as little research currently exists on this topic. Based on de Wit et al. (2019), it is believed that variation in gestural production may elicit different results of the previously mentioned constructs, as the chance may be higher that one of the gestures is recognised. As such, the following two research questions are central to this study:

"How do gestures in robot-assisted language learning affect children's second language word learning performance, engagement with the robot and their perceived anthropomorphism of the robot?"

"To what extent does the addition of variation in robotic gesture production affect a children's second language word learning performance, their engagement with the robot and their perceived anthropomorphism of the robot?"

2. Theoretical Framework

2.1 The Importance of Gestures in Language Development

The development of sophisticated language is something that sets humans apart from others in the animal kingdom. It has been essential to the way humans have evolved and formed communities, by allowing us to communicate. Yet a sizable portion of the way we communicate is not even verbally or written, but instead non-verbal behaviour: body movements, tone of voice and gestures. For human-human communication, gestures have been found to be integral to the way we communicate, increasing the understanding of messages, and allowing us to express additional information when used in conjunction with speech (Hostetter, 2011). Gestures may be best defined by Adam Kendon (2004, p. 18) as being an evident, specific movement of (a part of) the body, recognised as such by other partakers in an interaction as being a way of expressing meaning, emotion or thought. There are many ways to express these gestures, as suggested by McNeill (1992), who defined four main types: *iconic* (where the physical shape or way of enactment of the gesture alludes to the referent's shape or movement), *deictic* (a referential gesture depicting a spatial relation, i.e. pointing at something), *metaphorical* (a reference to a more abstract concept, such as a hand waving forward meaning 'future') and *beat* gestures (a mostly rhythmic type of gesture, carrying no particular message).

It has long been alluded that gestures are tightly linked with language learning, yet what mechanisms underlie this link is still widely debated. Some state that we have so-called 'mirror neurons' in our brains, small cells that can store information regarding certain physical actions seen by others. These mirror neurons facilitate *action understanding*: the process by which we create an internal description of an action that we can utilise in future behaviour (Rizzolatti, Fogassi, & Gallese, 2001). Rizzolatti and Craighero (2004) argue that these neurons laid the foundation for language, with speech forming out of gestural communication. However, there is strong opposition to this theory from those that believe that these mirror neurons have never been observed within humans (Hickok, 2009). Corballis (2003), proponent of the mirror neuron theory, does make a case for speech being formed out of gestural communication. He notes that humans often use gestures in sync with speech. Furthermore, he mentions that sign language used by deaf people has similar properties to spoken language; children who grow up using only sign language generally go through the same stages of language acquisition, oftentimes reaching certain stages earlier than their vocal peers.

For language development in children, gestures have been found to be important too. Children have been found to learn gestures before speech, generally starting with meaningless utterings combined with gestures (Butcher & Goldin-Meadow, 2010), before moving to word and (deictic) gesture combinations (Özçalişkan & Goldin-Meadow, 2005). After that, children start combining words themselves, yet the effects of gestures stay ever-present. Iverson and Goldin-Meadow (2005) touched upon this and found that two-word combinations can be predicted based on earlier gesture and word combinations made by a child. They suggest that these gestures may help with language acquisition, indicating that gestures may be a precursor to a willingness to learn specific verbal input. They give an example of a child pointing (a deictic gesture) at a hat that their father is wearing while uttering "dada". Upon which he responds by saying "That is daddy's hat", allowing the child to combine their own gesture with a newly learned word (Iverson & Goldin-Meadow, 2005, p. 370). This process may indicate a need for so-called *grounding* in young children.

Grounding is essentially a way of understanding the world through sensory experiences, and has been found important in language development, as language is often learned and applied using real world concepts (Matuszek, 2018). Words like "blue" or "big" are relatively abstract concepts by themselves, which are typically only understood when they are linked to words that have been established as a common ground between speaker and receiver. For example, a child may understand that there is a size difference between an elephant and a mouse. By linking the word "big" to the elephant and "tiny" to the mouse, the words are grounded based on their referential contexts. Gestures have been suggested to accommodate this effect, by allowing for that common ground to be established (H. H. Clark & Brennan, 1991). This may indicate that the gestures produced in Iverson and Goldin-Meadow (2005) satisfied a need for grounding in young children, by showing others that they understand a certain concept for which they do not yet have a word.

This gestural learning ability does not stop there, however. Young children have been found to more effectively learn verbs when presented with iconic gestures (Mumford & Kita, 2014). Even beyond that, children have been found to learn other things using gestures as well, such as mathematics (Cook, Duffy, & Fenn, 2013; Cook & Goldin-Meadow, 2006). Cook and Goldin-Meadow (2006) found that an instructor using gestures to teach a mathematical solution, led to children (nine- to ten-year olds) copying these gestures when they were asked to solve a mathematical problem on their own. The use of these gestures led to higher performance in post-test tasks, but interestingly also led to higher performance in mathematical tasks that had not been seen previously. In essence, the children were not merely mimicking the gestures, but were also capable of understanding the meaning of the gestures, and effectively reusing them in different situations.

Largely, it seems that gestures can provide a *scaffold* when learning, a temporary support strategy that helps learning by providing hints (Sawyer, 2014), and has been found to aid initial learning of new

information (Alibali & Nathan, 2007; McGregor, 2008). A review by Macedonia (2014) suggests that gestures can aid in L2 learning too; in a number of studies, gestures led to significantly better word retention when using congruent gestures, compared to merely a combination of hearing and reading. Other studies have shown that the use of deictic (e.g., Morett, Gibbs, & Macwhinney, 2012) and iconic gestures (e.g., Kelly, McDevitt, & Esch, 2009; Macedonia et al., 2011) are beneficial to L2 learning in adults. Further still, Macedonia and Knösche (2011) found that when teaching adults (fictitious) sentences using videos where humans performed gestures over the course of six days, gesture re-enactment (the act of 'imitating' gestures when asked) led to significantly better memory performance in the participants, compared to mere audio-visual repetition.

Similar results have been found for young children; Mavilidi et al. (2015) for example found a significant difference in recall of L2 words after having been taught the language using either a gesturing or non-gesturing condition with human tutors. The children (mean age of four years, eleven months) were able to more easily recall words in both free-recall (asking children which words they could still remember) and cued recall (showing an image and asking what the correct word for the image is) sessions when they were presented with gestures. Mavilidi et al. do however note that in all cases, the differences between test scores were rather small, arguing that this is likely caused due to L2 vocabulary learning inherently being a challenging task for young children.

Tellier (2008) found that children (mean age of five years, six months) were significantly better at remembering L2 words when they were presented with recordings of gestures performed by humans, instead of images. It should be noted that children were asked to re-enact the gestures and that sample sizes were rather small (n = 20). Tellier does however state that the addition of gestures may bolster memory, as it triggers both verbal and non-verbal modalities in the brain, as explained by the *dual coding theory*. This theory by J.M. Clark and Paivio (1991) explains that learning can improve when one is presented with a combination of multiple (non-)verbal modalities, building a richer pool of representations as more modalities are added. For example, when one is presented with a mere verbal representation of a word, this representation is stored once in the brain. However, a subsequent addition of gestures to represent that same word at the same time, allows it to get stored in a second, separate section of the brain linked to the first section. This additionally stored representation helps by building up a network of references the brain can call upon when memorising something.

2.2 Robots and Language Education

The previous studies show that language taught by humans using gestures can improve learning. Yet, recently, particular interest has been taken in the language tutoring capabilities of robots for children.

Largely, the reason for this interest stems from the advantages robots have compared to more traditional media such as computers. Belpaeme, Baxter, De Greeff et al. (2013) particularly noted a social robot's adaptability, with them being able to work in various educational and therapeutic settings. Specifically for RALL, robots have two distinct advantages (van den Berghe, Verhagen, Oudgenoeg-Paz, van der Ven, & Leseman, 2019). As previously mentioned, their physical nature is thought to be important to language learning, allowing for the handling of objects as well as bodily movements and gestures that can aid in language learning (Mavilidi et al., 2015; Rowe & Goldin-Meadow, 2009). Secondly, social robots often have a human-like appearance, which allows humans to anthropomorphise them: attributing them with human characteristics and behaviours (Bartneck, Kulić, Croft, & Zoghbi, 2009; Beran, Ramirez-Serrano, Kuzyk, Fior, & Nugent, 2011; Duffy, 2003).

Yet Belpaeme, Baxter, De Greeff et al. (2013) also note several challenges that have yet to be overcome. Ironically, many of these stem from technological hurdles, with current technology levels not being sufficiently high enough to accurately understand a child's unique form of speech (Kennedy et al., 2017) or the robot's perceptive system not being as advanced as that of a human (Bajcsy, Aloimonos, & Tsotsos, 2018), limiting the interaction possibilities. The result is that many current robotic studies rely on a Wizard of Oz method (Riek, 2012): children think they are interacting with a real self-actualised and autonomous agent, while in actuality the robot's actions are driven by a researcher behind a computer. In many ways this lowers ecological validity of these findings, as future robot products would act very differently in non-lab settings. Therefore, more research is needed where robots act autonomously and are present in the learning environment of children.

Finally, Belpaeme, Baxter, De Greeff et al. (2013) mention how difficult it is to measure the effectiveness of robots for children. Normally, there are various methods, scales, and questionnaires that an aid in understanding the perception of robots by humans. Yet these are often not viable for children; questions have to be trivialised for a young child to understand them, as they have been found to have difficulties with abstract Likert-scale questions (Mellor & Moore, 2014; also see Shields, Palermo, Powers, Grewe, & Smith, 2003). And with this trivialisation, accuracy of the measurements decreases rapidly.

2.2.1 Robots and Word Learning

In a review of 33 articles by van den Berghe, Verhagen et al. (2019), thirteen articles focused on robots and both L1 and L2 word learning, where various results were found. In general, it can be said that children were able to learn language together with robots, but not necessarily always better than with peers or adults. Interestingly, the number of words learned in many longitudinal studies was also quite low, generally only one or two words. This is in contrast to shorter studies that generally showed more distinctive results in number of words learned (e.g., de Wit et al., 2018; Kory Westlund, Jeong, et al., 2017; Tanaka & Matsuzoe,

2012). What was also apparent in the review was that generally, learning gain is the most used measurement to evaluate the effectiveness of robots. Most, these types of studies measured receptive vocabulary knowledge (similar to cued recall as mentioned by Mavilidi et al. (2015)), as opposed to productive knowledge (free-recall). The review also suggests that there were differences in age, with older children (ages nine and up) and adults generally being able to learn more quickly from robots than younger children, though this assumption was not specifically verified. Finally, van den Berghe, Verhagen et al. note that in general for all studies, participant counts were relatively low, typically between ten to forty participants.

When looking more specifically at the differences between robot and human tutors, a study by Kory Westlund et al. (2017) showed that a child's perception of non-verbal behaviour (gaze direction and body orientation) by a tutor was similar in both robotic and human tutors, with word learning performance being similar as well. Comparable results were found by Mazzoni and Benvenuti (2015). An L2TOR study[1] found that children tended to show similar learning performance with a robot tutor compared to a human tutor when both used iconic gestures.

L2TOR ('el tutor'; www.l2tor.eu) is a large-scale collaboration in robot-assisted language learning (RALL) set forth by several universities in Europe, on which the current study builds upon. This project attempted to gain a better understanding of how robots could aid in the acquisition of second languages for young children. In various studies, a social tutoring robot was used to teach children a new language using their own social and referential world. The robot employed various tactics, both verbal and non-verbal (in the form of gestures and body language), to aid in this endeavour. Together, these seem to suggest that robots can act as a suitable addition to a teacher's repertoire of tools assisting with teaching language to children.

2.2.2 Robotic Gestures and Word Learning

Curiously though, while extensive research has been done on the various types of effects gestures can have on language learning, as well as robots and language learning, little research exists on the combination of robotic gesture production and language learning. At present, there seem to be few robot studies that have looked at using gestures to teach children words.

A study by van Dijk, Torta and Cuijpers (2013) attempted to assess L1 verbal message retention in senior participants (mean age, 67 years), when accompanied by either a gesture or no movement. In a single session, a robot used gestures to portray subjects, verbs, objects and adverbs in the message. Results showed that gestures permitted significantly higher recall of verbs, with no differences found for the other word-types.

[1] This study is a yet unpublished study. The preliminary report can however be found on the L2TOR website, Deliverable 7.4, chapter 2, pages 7 – 15: http://www.l2tor.eu/effe/wp-content/uploads/2015/12/D7.4-Evaluation-report-storytelling-domain.pdf

Two other studies were both L2TOR studies. The first study, by Vogt et al. (2019), examined whether robots could effectively teach children several English words with the help of gestures in a longitudinal study. Over the course of six lessons (and one additional recap lesson), children were exposed to the English words through either a tablet and a robot employing iconic gestures, a tablet combined with a robot that did not produce gestures, or merely a tablet. With 194 participants (mean age five years, eight months), results showed that children learned more words in all experimental conditions compared to a control condition (where children were not exposed to any lessons). However, no differences were found between the conditions. In contrast to Vogt et al., de Wit et al. (2018) did find significant differences between conditions where a robot either employed iconic gestures or when it did not. In their study, the results of 61 children (mean age five years, two months) showed that on average, children performed slightly better on an immediate post-test, and much better on a one-week delayed post-test when the robot used gestures.

There were two notable differences between de Wit et al. and Vogt et al. The first difference stems from the duration of the study, with de Wit et al. taking only a single session. As seen earlier, single session robotic studies have shown higher learning gains than longitudinal studies (Gordon et al., 2016; Kanda, Hirano, Eaton, & Ishiguro, 2004; Movellan, Eckhardt, Virnes, & Rodriguez, 2009). Secondly, the words in Vogt et al. were more abstract than in de Wit et al. The study by Vogt et al. focussed teaching on spatial (e.g., 'in front of', 'climbing') and mathematical concepts (e.g. 'fewer', 'take away'), while de Wit et al. focused on more tangible words in the form of animals, which may have been easier to interpret or recognise by the children.

All in all, the previous studies showed that the use of gestures is ingrained in the (language) development of children. Robots too have been found capable of teaching a second language to children, though mixed results were seen when they employed gestures during teaching. However, as the current study is mostly based on the design of the study by de Wit et al. (2018), it seems plausible that robotic gesture use can lead to increased word learning performance. Therefore, the following hypothesis is posited:

H1. While teaching target words in a second language, using a robot to portray these words using iconic gestures will lead to an increase **(a)** in second language learning and **(b)** vocabulary retention, compared to a robot using no gestures.

2.3 Variation in Gestural Learning Stimuli

Interestingly, while much research has been done on gestures in general, there seems to be little research done on variations in semantic gesture production. There are however various ways to symbolise words when

using gestures (van Nispen, van de Sandt-Koenderman, Mol, & Krahmer, 2014). As seen in the dataset aggregated by de Wit et al. (2019), there are multiple ways to symbolise, say a pencil, using gestures. Some people use their arms and hands while stretching to make a pointy appearance, symbolising the pointy end of a pencil. Others still, attempt to portray a pencil by making a writing motion with their hands. Some like to add an additional reference to a long, straight object first before performing a second gesture. In the case of different words, some even used a pointed finger to 'trace' the outline of the object they attempted to portray. Müller (1998, as cited in Masson-Carro, Goudbeek, & Krahmer, 2016; Mittelberg & Evola, 2014) refers to these different types of iconic gestures as *representational gestures*, which are divided into four modes: *drawing* (using a finger to trace the silhouette of an object), *moulding* (using the hands to make a sculpt of an object, by forming a crown for example), *acting* or *imitating* (pretending to open a door) and *representing* or *portraying* (where the hands pretend to be an object, like a flat hand representing a piece of paper).

Further evidence of this can be seen in McNeill (1992). When asking different people to describe an event they had all witnessed, each tried to explain how a cartoon character tried to climb up a drainpipe. Interestingly, the participants all used different gestures accompanied by their explanation, yet all their gestures shared a common denominator: an upwards moving motion. McNeill found an explanation in that each participant had made their own choices in what they found salient in what they had seen. In essence, every person created their own mental imagery of what they had seen, and gestures helped them convey this particular imagery. It seems that the forming of these mental representations are influenced by personal experiences (Wyer, 2007), social interactions (Levine & Resnick, 1993) or even cultural backgrounds. Kita (2009), for example, found that conventionalised gesture use varies greatly between different cultures. Using the index finger and thumb to form a ring means "OK" in many European cultures, yet some cultures differ from this interpretation. In France for example, this gesture can mean "zero", while in Greece this can instead mean a bodily orifice. What the studies by McNeill and Kita show is that gestures allow for the communication of personal, mental concepts. But this communication is also inextricably influenced by a predetermined agreement on a gesture's meaning. This may give an indication of how different interpretations of gestures can form, and why there are differences in what individual people find the most effective way to communicate their own definition of a word through gestures.

Like language has different synonyms for the same word, it seems that gestures too can have different productions explaining the same concept. It may thus be possible that this phenomenon works both ways; correct interpretation of gestures may rely on what individually available references to an object exist within a person. This notion is supported by Piaget's *Theory of Cognitive Development* (Piaget, 1952; Ültanır, 2012), which states that a child's brain structures knowledge in so-called *schemata*, a way organising and grouping knowledge. For example, despite their visual differences, both German Shepherds and Golden Retrievers

share certain characteristics (e.g. both races have four legs and paws, snouts, panting behaviours, etc.) causing them to both conform to a 'dog' schema. As children learn and experience new things, new characteristics are added to existing schemata through a process called *assimilation*, further developing these schemata. It may be possible that especially young children have not yet had the chance to assimilate many different characteristics, leading to lesser established schemata of certain concepts. Being able to target a single concept in multiple ways (by using variations in gestures, for example) may lead to a higher probability that the child recognises a concept that is present in one of their existing schemata. While this theory has often been put in contrast to the previously mentioned dual-coding theory (J. M. Clark & Paivio, 1991), both share a similar outcome: knowledge increases as more and different types of information are combined.

Different areas of research also understand this, for example in the education area, there exists the Theory of Variation (Marton & Booth, 1997). This theory states that each person has a different understanding of the world based on their pre-existing knowledge. When they are presented with new information (via teaching) that does not conform to concepts and beliefs of the world that have already formed within a person, they may resist this new information which inhibits the learning process. In order to accurately tap into the variations in understandings, teachers need to use variations in stimuli to help the learner understand what is and is not part of the so-called *object of learning*. In other words, the amount of knowledge one has is limited by how many variations of possible outcomes he has learned. For younger children, the time they have had to learn many different variations of existing knowledge is limited, and thus it is possible that using multiple gesture variations is more likely of triggering one of the pre-existing notions a child has with a certain concept.

Altogether, while there is evidence in developmental research that variation can aid learning, little is known about how this translates to gesture variations and their effects on second language learning. Combining these theories, it seems plausible that using variations in gestures can aid in developing more robust schemata. The higher number of gestures allows for more opportunities for a child to recognise a gesture based on their personal, pre-existing knowledge. Based on these theories, the following hypothesis is posited:

> **H2.** While teaching target words in a second language, using a robot to portray these words with a new, previously unused gesture variation for each target word, will lead to an increase in **(a)** second language learning and **(b)** vocabulary retention, compared to using only static gestures or no gestures.

2.4 Robot Engagement through Gestures

Gestures and their varied counterparts may also lead to an additional benefit indirectly related to learning: it can make the robot more interesting. In the beginning phases of exploring the opportunities of robotic design, most studies focused on solving technical issues with robots. However, as these technical hurdles are slowly overcome, the attention shifts to other areas in HRI. A similar shift was seen in human-computer interaction (HCI, of which HRI is a sub-domain) where after focusing on mere usability for years, more and more focus was put on making a more enthralling experience (Hassenzahl & Tractinsky, 2006). O'Brien and Toms (2008) state that in HCI, various user experience elements have been hypothesised to be linked to a change in user engagement. They posited the following definition of engagement: "Engagement is a quality of user experiences with technology that is characterised by challenge, aesthetic and sensory appeal, feedback, novelty, interactivity, perceived control and time, awareness, motivation, interest, and affect." (O'Brien & Toms, 2008, p. 949). According to O'Brien and Toms, engagement can lead to higher attention, (intrinsic) motivation and curiosity, which in HCI subsequently has been linked to an increase in learning (Huizenga, Admiraal, Akkerman, & Ten Dam, 2009; Liu, Horton, Olmanson, & Toprac, 2011).

There are several studies that link a robot's motion in general to increased engagement. They have shown that the use of body and head gestures can lead to increased attention towards the robot (e.g. Michalowski, Sabanovic, & Simmons, 2006; Sidner, Kidd, Lee, & Lesh, 2004; Sidner, Lee, Kidd, Lesh, & Rich, 2005). Furthermore, motion in robots also seems to affect social engagement, defined as "the process by which two (or more) participants establish, maintain and end their perceived connection during interactions which they jointly undertake" (Sidner et al., 2004, p. 1).

A study by Burns, Jeon, and Park (2018) found that participants showed significantly more positive (facial) emotions and engagement when they interacted with a robot that mimicked their own movements. This further extended to mood contagion (switching and matching the mood of a different social agent), which occurred more often when the participants were in an experimental group than those that were in a control group. The experimental group was more likely to imitate a wider selection of emotions portrayed by the robot, indicating that robotic motion can have substantial effects on the cognitive processes of humans. In an observational study by Sabanovic, Michalowski and Simmons (2006) a robotic receptionist was used over the course of two days to see how social interactions form in a natural environment. In their interactions, people could ask for directions or the weather, upon which the robot reacted. Results showed that the robots elicited more social interactions with humans when they portrayed either spatial- or gestural movement, suggesting that a robot in motion is perceived as more sociable.

It should be noted here however that all previously mentioned studies either did not report on the ages of the participants (Michalowski et al., 2006; Sabanovic et al., 2006; Sidner et al., 2005) or were adults

(Burns et al., 2018). Therefore, it is unknown if and how these effects carry over to interactions with young children, as it may be entirely possible that the perception of social characteristics of robots is different for them.

Evidence has also been found that gestures can actually decrease engagement. Huang and Mutlu (2013) for example found that metaphoric gestures in a robot interaction (gestures that represent a metaphor, like a hand waving forward meaning "future") decreased engagement in both male and female adults. Huang and Mutlu speculated, however, that this was largely caused by the amount and the abstract nature of the metaphoric gestures in their study, possibly causing distraction in the participants.

Interestingly, there seem to exist few robot studies that pertain to the effects of a robot's motion on engagement in education. De Wit et al. (2018) found that robotic gesture use, positively contributed to engagement for children. However, this study took a rather general look at engagement, without making further digression in what it means for robot-learner interaction. Furthermore, it used a random selection of participants to assess engagement in child-robot interaction via a single Likert scale question based on two short, five-second video clips per child. This does beg the question how accurate and valid this assessment of engagement is.

On the whole, it does seem likely that salient task-related motion can contribute to higher engagement, hence the following hypothesis is posited:

H3. Children who interact with a robot portraying gestures will be more engaged than children who interact with a robot which does not portray gestures.

Engagement through variation

Curiously, there is limited research on how engagement changes when a robot performs monotonous, identical motion during an interaction. This is notable, as robots are often associated with being repetitive (Haring, Mougenot, Ono, & Watanabe, 2014; Ray, Mondada, & Siegwart, 2008). One would expect there to be more research on robots that break monotony, by alternating between different responses and see how it affects robot-related engagement, and the kinds of reactions it elicits. Of what little research there is, Tanaka, Cicourel and Movellan (2007) found that toddlers showed increased interest in robots that showed a wider breadth of behavioural patterns, than when it showed repetitive patterns. Moreover, Ros, Baroni and Demiris (2014) found that over the course of multiple repeated interactions, the novelty effect wavered for children, suggesting that repetitive behaviour of a robot leads to lower motivation. Finally, Belpaeme, Baxter, Read, et al. (2013) mentioned that repetitive usage of verbal language in robot interactions can be tiresome. According to them, adding variation helps prevent issues in engagement with robots and is key for

maintaining this engagement. However, Belpaeme, Baxter, Read, et al. provides no sources as to why this happens. Together, these studies outline that less repetition can lead to higher engagement. Nevertheless, there are several things to note about these studies. Tanaka et al. (2007) focused only on longitudinal effects (45 sessions) of motion and provided no insights in single session engagement. Ros et al. (2014) did look at single session engagement but had no control condition to compare it to. On top of that, both studies recruited fewer than twelve children. Finally, Belpaeme, Baxter, Read et al. (2013) made merely a notion, and provided no evidence.

While the previous studies do not provide a robust insight in what the effects of variation in robot interactions can elicit, it does however seem plausible that a robot portraying variation in gestures may increase engagement. Therefore, the following hypothesis is posited:

H4. Children who interact with a robot portraying multiple gesture variations will be more engaged than children who interact with a robot using only single or no gestures.

2.5 Robot Anthropomorphism through Gestures

Finally, gestures may also affect how children anthropomorphise the robot. As previously mentioned, anthropomorphism is the tendency of humans to attribute human-like characteristics to robots. If a robot is seen as exhibiting a high number of anthropomorphic characteristics, it is regarded as being more humanlike. This plays an especially large role when attributing social features to robots, like eyes or a mouth which indicate a social capability in the robot (Duffy, 2003). Duffy notes that these anthropomorphic characteristics are vital for establishing meaningful interactions with humans, stating that more anthropomorphic features can lead to a stronger expectation of a system's performance by users. However, Duffy also noted that effective use of anthropomorphism is a delicate balance. A social robot exhibiting too many human-like features or being too intelligent may lead to perceptions of selfishness or weaknesses similar to humans, negating its reason of existence. If a robot exhibits certain affordances (properties that indicate a possible use or action of an object), say a mouth indicating that it can speak, then certain expectations are set by users. Not meeting these expectations, such as a robot that never speaks even though it has a mouth, can then lead to confusion and subsequently lower perceived anthropomorphism (Bartneck & Forlizzi, 2004).

Anthropomorphism then, should be regarded as an important feature in the designing of future social robots. Most of these anthropomorphic characteristics seem to stem from a robot's physical 'surface look' (its gender or eyelashes for example) and its 'body-manipulators' (torso, hands and arms), according to Phillips, Zhao, Ullman and Malle (2018). But anthropomorphism goes beyond mere physical attributes.

Behaviour too is an important anthropomorphic indicator. A study by Tung (2016) showed that children find robots more physically and socially attractive when they exhibit social cues such as facial expressions, speech and gestures.

Fink (2012) on the other hand, makes a further discernment between the anthropomorphic design of the robot and anthropomorphism itself. In many ways, anthropomorphic design of the robot is influenced by the robot's behaviour, communication and its physical attributes, the aforementioned affordances. According to Fink, these can be steered and adjusted to elicit certain reactions. The reaction on this design is anthropomorphism itself, in essence a social perspective that develops as a response to interacting with the robot based on the perceived emotions, motivations, and intentions. Furthermore, Epley, Waytz, Akalis and Cacioppo (2008) suggest that anthropomorphism is inherently affected by individual, cultural, environmental and situational aspects as well. Lemaignan, Fink and Dillenbourg (2014) further explain that anthropomorphism of a robot can increase as it exhibits disruptive behaviour, a sudden change in the expected behaviour by a robot. This is especially interesting for the current study, as the inclusion of additional gesture variations may influence the perception of anthropomorphism by children, as the children would be introduced to a new way the robot moves each time.

Gestures themselves too may have an effect on the perception of anthropomorphism. Salem, Eyssel, Rohlfing, Kopp and Joublin (2013) attempted to understand whether non-verbal communication had an effect on the perceived anthropomorphism of robots, as well as its likeability. Indeed so, adults perceived the robot as being more humanlike, as well as being more likeable if it exhibited gestures. This effect was even more salient if the gestures were incongruent with its speech, further increasing anthropomorphic perception, but at the cost of task-performance.

Anthropomorphism has a role to play in digital learning as well, as it can increase learning performance and intrinsic motivation of students (Schneider, Häßler, Habermeyer, Beege, & Rey, 2019). For second language learning, van den Berghe, de Haas et al. (2019) found correlations between word learning and anthropomorphisation of robots by children. On average, anthropomorphism stayed level between pre- and post-test. However, a large variability was visible when looking at the children individually, with them showing vastly different scores at the two test points. These differences were found to correlate with word learning; if a change was positive (i.e., the robot was perceived as more human-like between pre- and post-tests), then the child was more likely to show higher learning gains. Vice versa, a similar pattern was visible. Even though these correlations were weak, van den Berghe, de Haas et al. suggests that this may be related to the child's expectations. If the robot exceeded the child's prior expectations, this may have increased their engagement during the game and subsequently increased their word learning compared to a robot that failed to live up to their expectations.

There is still uncertainty how this pertains to young children, as there seem to be few resources dedicated to testing anthropomorphism in this target group. The Godspeed questionnaire (Bartneck et al., 2009) is one of the most widely used anthropomorphism questionnaires in the robot field. However, as previously mentioned, it uses Likert scale questions, which provides challenges when used with young children. At present, the only questionnaire that seems to exist for examining robot anthropomorphism for young children is the one used by van den Berghe, de Haas et al., which was based on a questionnaire for living and non-living objects by Jipson and Gelman (2007), where one of the included objects was a robotic dog.

In conclusion, it is likely that children playing with a robot that uses gestures anthropomorphise the robot more than those that play with a robot that does not use gestures. However, as examining the anthropomorphism of robots by young children is still relatively new and fairly difficult to do consistently, especially without any tried-and-true testing methods for this target group, this part of the study will be explorative in nature. Therefore, the following research question is set out:

> ***RQ1:*** *To what extent do gestures influence the perceived anthropomorphism of the robot by children in an educational setting?*

3. Method

In large, the current study is based on the design of the study by de Wit et al. (2018), containing several new aspects, most notably the introduction of variation in the robot's iconic gestural production. In both the current study and the study of de Wit et al. a robot was used to teach a second language supported by the use of iconic gestures. However, while de Wit et al. focused on gestures in combination with an adaptive system to tailor the learning experience to the user, the current study forwent this system and instead focused more directly on looking at the effects of gestures on learning performance, engagement, and perceived anthropomorphism. Furthermore, while in the original study gestures were composed and hand-edited by a single designer, the current study sourced the gestures from a database of crowdsourced gestures from de Wit et al. (2019). Finally, several changes were made in the design of the study, including adding additional participants, additional measurement constructs and updates in the pre- and post-testing measures to further improve the robustness of this study.

3.1 Design

In order to test whether variations in robotic gesture production can aid in second language learning, a three-group between-subjects experimental design was set up. All three conditions featured a robot with which the children interacted to learn six English words. In the first condition, the no-gesture condition, the robot used a static stance and produced no gestures. In the second condition, the single-gesture condition, the robot produced a single, repeated gesture each time for each unique target word that was presented. In the final condition, the varied-gestures condition, the robot varied its gestures to show a new gesture each time a target word was presented. As dependent variables, measurements were made on the children's word learning gains and retention, and the engagement during the experiment. Furthermore, perceived anthropomorphism was used as an exploratory measure both before and after the experiment. Based on the previous literature, four hypotheses were posited. Finally, the present study was approved by the Ethics Review Board of Tilburg University, and was pre-registered to prevent confirmation biases and wrongful interpretations of the data. An overview of the analysis plan can be found in the preregistration document available at http://aspredicted.org/blind.php?x=af7es6.

3.2 Participants

In total, 116 children participated in this study (with nearly 200 admissions in total). However, as set out in the pre-registration, there were several exclusion criteria: Only native Dutch children were allowed to participate in the study, with Dutch as their native language. Bilingual children were excluded as previous

research has shown that bilingual children are more apt at learning an additional language (Cenoz, 2003). Furthermore, children who knew fewer than five words during the Dutch pre-test, or more than four of the words during the English pre-test were also excluded. Finally, if a participant missed any one of the measurements, they were also excluded from the final dataset.

Therefore, 22 children were excluded from the analysis. The reasons vary from technical or procedural issues (12), bilingualism (3), knowing more than four English words in the pre-test (3), and finally incomplete results due to attrition (4). This resulted in a final tally of 94 participants that were included in the analysis. The children had an average age of 5 years and 3 months (*SD* = 9 months) and were randomly assigned to one of the three conditions: no-gesture condition (*n* = 33), single-gesture condition (*n* = 32) and the varied-gestures condition (*n* = 29), whilst maintaining a balance in age and gender (Table 1).

Table 1. Total participant distributions per condition

	No-gesture	Single-gesture	Varied-gestures	**Total**
Participants				
n	33	32	29	**94**
% girls	48.5	43.8	58.6	**50.0**
% boys	51.5	56.3	41.4	**50.0**
Age in Months				
M	5 years, 3 months (*SD* = 9 months)	5 years, 2 months (*SD* = 9 months)	5 years, 4 months (*SD* = 8 months)	**5 years, 3 months (*SD* = 9 months)**

All children were recruited from Dutch primary schools, by contacting schools and providing them with an information letter accompanied by a consent form for parents (Appendix B). The information letter contained general information regarding the experiment, estimated duration of the study as well as references to the previous L2TOR studies (e.g., the video available on the L2TOR website).

3.3 Materials

The aim of the study was to teach children six words in English (namely: bridge, horse, pencil, spoon, stairs and turtle). These specific target words were chosen as they allowed for targeting of multiple gestural modes (as proposed by the earlier mentioned study by Müller (1998, as cited in Masson-Carro, Goudbeek, & Krahmer, 2016; Mittelberg & Evola, 2014) that could convey the word's meaning, while remaining uniquely differentiable from each other in both phonetical and gestural interpretation. Furthermore, care

was taken to make sure that the Dutch translations of these words did not sound too similar to their English counterparts, providing further reliability to the measurements.

A Softbanks Robotics NAO was used as a robotic tutor for the children. This robot stands 574mm tall and is capable of fairly human-like movements, however, it does have certain limitations. For example, the robot is capable of 96 degrees of side-to-side arm movement (18 degrees inwards, 74 degrees outwards; SoftBank Robotics (n.d.)) compared to a human's 200 degree side-to-side arm movement (50 degrees inwards, 150 degrees outwards; Washington State Department of Social & Health Services (2014)). Furthermore, while the robot does have three fingers, they are incapable of moving independently and can only be extended or retracted all at once. The limited range of movement led to issues with one-to-one mapping of the gestures sourced from the dataset produced by de Wit et al. (2019). Furthermore, additional movements made by the participants that would normally be filtered out when viewing humans performing them, were becoming significantly more salient when viewing the robot performing these gestures. The resulting gestures were difficult to understand even when one knew what word was being portrayed.

In the end a choice was made to use the dataset as a foundation of the gestures, and they were manually recreated and adapted to fit the dexterity of the robot. This led to the creation of 30 different gestures, five for each word. A perception study of this gesture-set was held amongst 19 adult participants (recruited via convenience sampling) who were presented with video recordings of all gestures via an online questionnaire. Per gesture they were asked which of the six target words they felt best matched the gesture, as well as how sure they were of their answers. The findings revealed that eight gestures scored poorly (scoring an agreement rate lower than 60 percent), nine scored moderately (between 60 and 70 percent) and thirteen scored strongly (above 70 percent agreement). The eight gestures that scored poorly were redesigned with the rest of the gestures being left as-is, with them being regarded as satisfactory when they would be combined with an auditory word during the experiment itself. Furthermore, the highest scoring gesture per word in the perception study was used as the gesture for the same word in the single-gesture condition. A complete overview of all gestures can be viewed in Appendix A.

The setup was completed with a tablet that allowed the child to indirectly play a game with the robot, a computer that served as an interaction device for the pre- and post-tests, as well as two cameras recording the children's interaction from a side and a front vantage point (Figure 2). The children themselves sat at a table with a chair directly in front of the robot, with the tablet residing in between them.

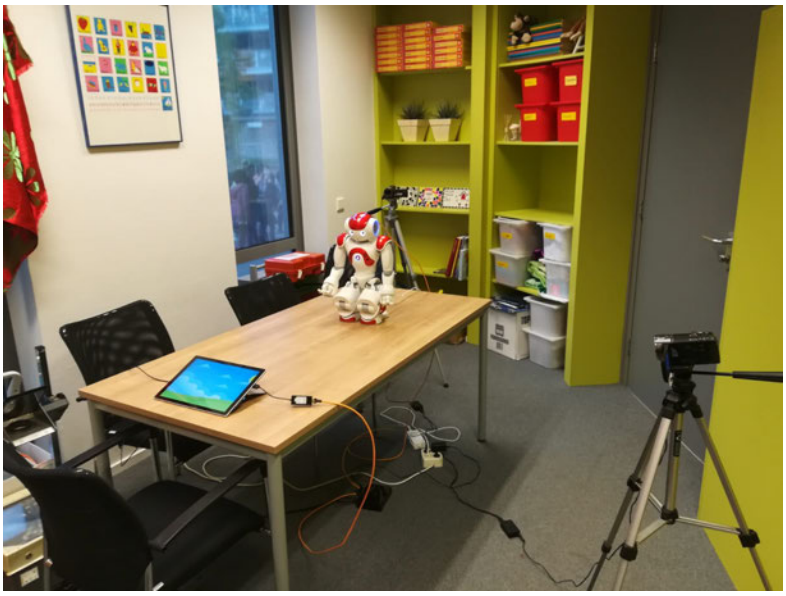


Figure 2. General setup of the experiment. Note: setup sometimes differed slightly per location due to room layout.

3.4 Measurements

3.4.1 Vocabulary Knowledge and Retention

In order to test for vocabulary knowledge and retention, a comparison was made on test-scores recorded at three separate times: before the experiment (pre-test), immediately after the experiment (immediate post-test) and a final time approximately one week after the experiment (delayed post-test). In order to alleviate additional testing fatigue during the experiment, the pre-test was taken a minimum of one day before the experiment.

The pre-test was conducted to ascertain the child's prior knowledge (both their Dutch (L1) knowledge and their English (L2) knowledge) of the chosen target words. This test was conducted on a laptop, where the child was shown images of all six target words positioned randomly on the screen (Figure 3). A native speaker pronounced one of the six words, before asking the child to click on the matching image. This was done for all six words, first in Dutch, then in English. To improve reliability and robustness of the earlier study done by de Wit et al. (2018), each target word in the English pre-test was tested three times (compared to one round in de Wit et al., and many other robot studies as mentioned earlier by van

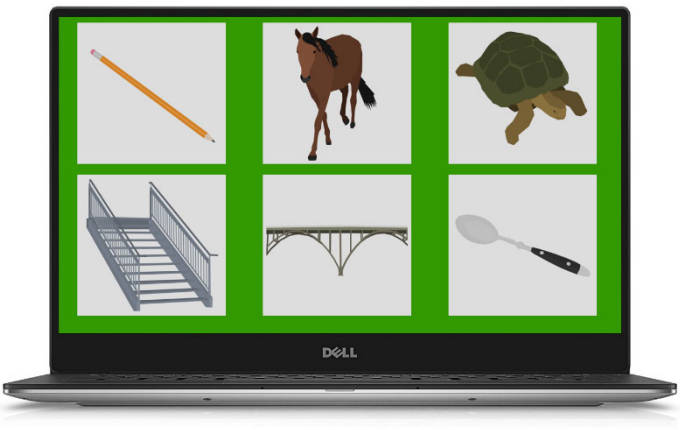


Figure 3. Laptop interface, with six images corresponding to the target words, that was used in the pre-test and post-tests.

den Berghe, Verhagen et al. (2019)). The English pre-test used three different sets of images (Figure 4a, b and c): once with a cartoony illustration of the concept (identical to the one that would be used in the experiment), once with a photo of the concept and once with a line-drawing of the concept. Three different images were chosen as to make sure the child correctly understood the concept, instead of merely relying on recognising colours or composition of the image, aiding them in a potential guessing attempt.

If the child chose the correct image two out of three times, it was assumed that the child understood the concept correctly. The children had to have at least five words correct during the Dutch pre-tests, as well as a maximum of four words correct during the English pre-test in order to be eligible for inclusion in the final dataset. The Dutch pre-test only took one round using the cartoony illustrations, as it was expected that all children would correctly understand the words. For both the immediate post-test and the delayed post-test an identical testing setup was used.

Figure 4. Examples of images used in pre- and post-test showing a cartoon illustration of the concept (a), a photo of the concept (b) and a line drawing of the concept (c).

3.4.2 Engagement

During the experiment, two cameras recorded the entire interaction the child had with the robot. For analysis, these videos were trimmed to two, two-minute sections taking place in the fourth and twenty-fourth round of the experiment (n = 188). The fourth round was chosen as to allow the children to get used to playing the game in the training, whilst still obtaining initial engagement results. The twenty-fourth round was chosen in order to obtain near end-state engagement, before the child had been reminded that the training was nearly done. This selection is based on the study by de Wit et al. (2018), except that in the present study the video segments over which engagement was encoded, was increased from five seconds to two minutes. Initially, these were then coded by two independent raters on four different constructs: robot-child engagement, task-child engagement, valence, and arousal. These scales were based on two earlier robot studies. An overview of these codebooks can be found in Appendix C.

The study by de Haas et al. (2019) focused on measuring general engagement between the robot and child, and the task and child and was scored on a nine-point ordinal scale [-2, +2], with half increments. These scales focused on motivation, mental activity, satisfaction, and a need for exploration. Both separately measured engagement on two distinctive points. First, robot-child engagement indicated the level of focus in regard to the robot. A child ignoring the robot completely, showing no signs of investigating the robot and passively interacting with it, was regarded as low engagement. On the opposite side of the spectrum, if a child was actively looking or even talking to the robot, having uninterrupted focus on the robot, this was regarded as high engagement. Secondly, task-child engagement indicated the level of focus on the task at hand with the tablet. Low engagement here, indicated that the child was not paying attention when asked to perform a task, or focusing on other environmental factors in the room. High engagement was regarded as showing an active use of mental capacity to complete the task, and a driven, unbreakable level of concentration. It is thus possible for a child to achieve high task-child engagement, yet low task-child engagement and vice versa, depending on how pre-occupied they are with either the tablet or robot.

As it was felt that these scales did not completely capture the range of visible reactions by the children, three additional scales were added based upon the study by Rudovic, Lee, Mascarell-Maricic, Schuller and Picard (2017). Valence allowed for the recognition of feelings ranging from unpleasant (such as unhappy or dissatisfied reactions) and disappointment, to signs of happiness and joy. This scale was scored on a nine-point ordinal scale [-2, +2] with half increments. The second scale, arousal, allowed for the recognition of the level of excitement within children, ranging from sleepy and bored to active and responsive, again scored on a nine-point ordinal scale [-2, +2] with half increments. A -2 or a +2 on a scale referred to a negative or positive expression of the construct, while the levels in between are regarded as a lower intensity of either extremities. A 0 was regarded as a neutral state where the child shows no discernible emotions (valence) or limited physical activity (arousal). Raters were instructed to not base their decisions on instinct regarding the children's internal states, but instead focus on the visible manifestations of each construct.

Before coding all the videos, two raters were presented with a smaller sub-sample of videos (n = 50) and were asked to code these based on the available codebook. Over several discussions, the codebook was adapted. An analysis of the final coding showed that the second rater generally used less-conservative ratings than the first rater, even after discussing multiple coding sessions. As can be seen in Figure 5 (showing difference of scores between raters on child-task engagement), the second rater commonly scored 0.5 and 1 higher than the first rater. Similar patterns were found for the other constructs. As scores typically only deviated by these smaller increments, a quadratic weighted Kappa (a Kappa that allows for partial agreement, while putting exponentially more weight on larger disagreements) was used to determine the inter-rater

reliability. Recently, discussion has taken place on how to interpret the Kappa, with newer approaches (e.g., McHugh, 2012) being much more conservative than the original interpretations (e.g., Landis & Koch, 1977). Based on McHugh (2012), inter-rater agreements were found to be weak for child-robot engagement (κ = .58, p < .001), child-task engagement (κ = .50, p < .001), valence (κ = .50, p < .001) and arousal (κ = .50, p < .001). Partially, this seems to be explained by the fact that the first rater had seen more exorbitant behaviour of the participants during the actual experiment, behaviour that was often not visible in the video recordings. The first rater thus used a more conservative approach when coding the videos. In the end, the decision was made to only include the ratings of the first rater in the analysis in order to retain a more consistent dataset. Caution should be used when interpreting the results however, as this may have introduced a bias.

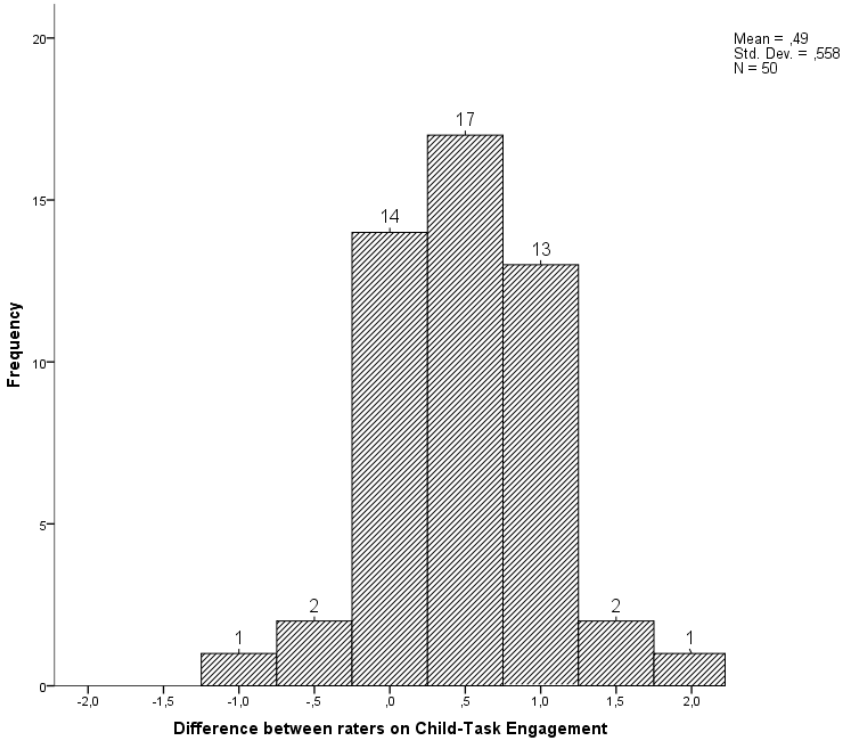


Figure 5. Graph showing differences between raters on the child-task engagement construct.

3.4.3 Perceived Anthropomorphism

In addition to the vocabulary knowledge pre-test, a pre-test questionnaire was held to ascertain the child's perception of their anthropomorphism of the robot. The questionnaire featured an adapted and updated version of the questionnaire by van den Berghe, de Haas et al. (2019). It consisted of thirteen questions that could be answered with 'yes', 'no' or 'I don't know', as well as featuring an open-ended component on why the child gave a particular response. In order to the reduce the length of the original questionnaire by de Haas et al., several questions were removed as they were not relevant to the present study. The remaining eleven questions were divided into two main categories, which tested for both a mental and a biological

component of the child's perceived anthropomorphism. Two additional questions were added based on the study by Kanda, Koizumi and Shimada (2012), namely "Would you like to play 'I spy with my little eye' (again) with Robin?" and "Would you like to learn other things by playing games with Robin?". An overview of these questions can be found in Table 2.

For each 'yes' answer, a single point was awarded, and no points were awarded for a 'no' or 'I don't know' answer. The two 'other' questions were not added to the total anthropomorphism score. In total, a maximum anthropomorphism score of eleven could be had, with a higher score meaning that the child attributed more human-like characteristics to the robot. A Cronbach's alpha showed that the internal consistency was acceptable for both the pre-test (α = .73) and the post-test (α = .77) questionnaires.

Table 2. Overview of the questions used to measure anthropomorphism in children. Adapted from van den Berghe, de Haas et al. (2019) and Kanda, Koizumi and Shimada (2012).

Do you think that Robin the robot…		
Biological	*Mental*	*Other*
…can feel it if you tickle him?	…can be happy?	Would you like to play 'I spy with my little eye' (again) with Robin?
…can feel pain?	…can be sad?	
…can see things?	…can remember something?	Would you like to learn other things by playing games with Robin?
…grows?	…knows a lot?	
…needs food?	…is smart?	
	…understand when you say something?	

3.5 Procedure

To ensure that all experimenters followed the same testing procedure, a protocol was written and distributed amongst all experimenters, which can be found in Appendix D.

To improve ecological validity of this study, all robot interactions took place in a classroom(-like) environment on the schools the children attended. Prior to the main experiment, all participants were introduced to the robot during a group introduction. Based on previous studies by de Wit et al. (2018) and Vogt et al. (2017), it has been found that introducing the robot helps alleviate anxiety issues during face-to-face interactions with the robot later on. Furthermore, framing the robot as a social entity rather than a mechanical entity has also been shown to attract more attention to the robot (Westlund, Martinez, Archie, Das, & Breazeal, 2016). The group introduction was followed by the previously mentioned pre-tests, though not necessarily on the same day. The group introduction took approximately 15 minutes, with the combined pre-tests taking approximately 10 minutes per child.

A minimum of one day after the pre-test, the children participated in the experiment which took approximately 25 to 45 minutes per child overall. First, the children were presented with a concept binding session. The child went through each target word, using a laptop, to expose them to the correct mappings between the target words and the portrayed concept. This mapping helped the children with the initial rounds of the experiment and prevented from turning the training into a guessing game. Furthermore, this helped establish a baseline for the children of what the correct answers are, instead of relying on their answers in the pre-test to be the correct ones. For each word, the corresponding image was presented on the laptop accompanied by a recording of a native speaker saying "Look, this is a [target in L2]. Do you see the [target in L2]? Click on the [target in L2]!".

The children were then divided into the three conditions: no-gesture condition, single-gesture condition, or the varied-gestures condition. The children then played thirty rounds (plus an additional Dutch and English practice round) of the game *I spy with my little eye* with the robot. During this game, the robot mentioned that he saw one of the six target words and announced, "I spy with my little eye, [target in L2]". Based on which condition the child was assigned to, the robot accompanied its announcement with either a gesture or none at all. As the game progressed further, the target words were each repeated for a total of five times. In the single-gesture condition, each unique word was accompanied with the same gesture, while in the varied-gestures condition, each subsequent word was portrayed by a new, previously unused gesture. The aim of this new gesture is to trigger a different interpretation of the word in the child's semantic network. The order of these gestures was randomised for each child.

After the robot's announcement of the target word, the child was presented with three images on the tablet: one correct image, and two distractor images (Figure 6). The child was asked to pick the image that matched the target word mentioned by the robot. To let the child know whether they were right or not, both the tablet and the robot provided them with feedback. The tablet highlighted the selected image and provided either a green, happy smiley for a correct answer or a red, sad smiley for an incorrect one. Simultaneously, the robot provided verbal feedback to let the child know they were either right or wrong. If the child entered an incorrect answer, they had to do an additional "repair" round where the robot restated the Dutch word ("You pressed [wrong answer], but I saw a [correct answer in Dutch].") followed by a

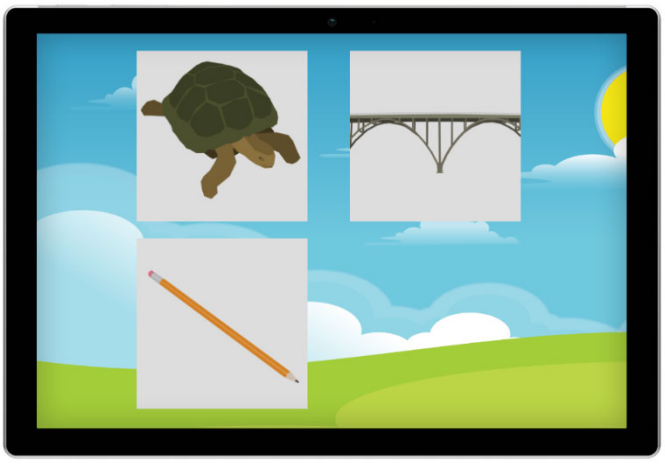


Figure 6. Tablet interface with distractor images.

request for the correct answer in English. To make it easier for the child, this repair round featured only one additional distractor (instead of two in the original round). In the varied-gestures condition, the same gesture was repeated as in the original round. This process was repeated for a total of thirty times, before the robot mentioned he enjoyed the game and said goodbye.

Immediately after the experiment, children were presented once more with the anthropomorphism questionnaire, followed by an immediate vocabulary post-test to assess whether they had learned any new words, after which they were allowed to return to their class. A minimum of one week after the experiment, children were given a delayed vocabulary post-test a final time, this time without an anthropomorphism questionnaire. This final session took approximately 3 minutes per child.

Finally, both the children and the schools received an appreciative gesture in the form of a papercraft model of the NAO robot for the children (Appendix G), and a box of chocolates for the schools. In addition to this, schools and parents who requested so, received a simplified version of the results attained in the current study in the form of an infographic (Appendix F). Furthermore, any children whose parents had filled in the consent form but were unable to participate in the experiment, were given a shortened group-version of the robot-interaction. No data was collected from these children.

4. Results

In the following results section, the pre-registered results will be analysed. However, as observations indicated that there may have been additional effects related to the other variables, an additional exploratory results section has been added containing further analyses. As these analyses have not been pre-registered, they should be interpreted with caution.

4.1 Descriptive Analyses

A one-way ANOVA confirmed that the children between the three experimental groups were similar in age, $F(2, 91) = 0.54$, $p = .582$. Similarly, a Pearson Chi-Square showed that there was no statistical difference in gender distribution between the experimental groups, $\chi^2 (2) = 1.39$, $p = .498$.

Normality was violated in several cases: Dutch Pre-test scores ($z_{Skewness} = -11.23$; $z_{Kurtosis} = 12.02$), English pre-test scores ($z_{Kurtosis} = 3.61$) and the immediate post-test score ($z_{Skewness} = 2.18$). For both the mental component of the anthropomorphism questionnaire normality was violated for the pre-test ($z_{Skewness} = -4.10$) and post-test ($z_{Skewness} = -4.95$), as well as the biological component of the questionnaire in both pre-test ($z_{Kurtosis} = -2.18$) and post-test ($z_{Kurtosis} = -2.56$). While these normality issues are expected, the results should be interpreted with some caution even though an ANOVA is fairly robust against this violation. For other analyses, the 95% bootstrapped confidence interval will be reported (5000 iterations).

4.2 Main Pre-registered Results

4.2.1 Word Learning

In order to test for word learning, a one-way repeated-measures ANOVA was performed, using condition and time as independent factors, and testing scores as dependent factors. The ANOVA revealed a main effect for second language learning over time, $F(2, 182) = 45.696$, $p < .001$, $\eta_p^2 = .334$. Post-Hoc tests using the Bonferroni correction revealed a significant effect between the English pre-test ($M = 1.32$, $SD = 1.08$) and the immediate post-test ($M = 2.44$, $SD = 1.88$), $M_{dif} = 1.10$, $p < .001$, and the delayed post-test ($M = 2.73$, $SD = 1.68$), $M_{dif} = 1.40$, $p < .001$. However, no significant difference was found between the immediate post-test and the delayed post-test ($M_{dif} = 0.30$, $p = .088$). It can be concluded that children learned from the training interaction with the robot, as well as retained the words in memory a minimum of one week later. The ANOVA did not reveal an interaction effect between word learning and the three conditions, $F(4, 182) = 1.58$, $p = .180$. Thus, no effect was found for gestures performing better than no gestures. Overall, while children did learn from the training in general, using gestures to teach them English words did not

make a difference for both word learning and retention. An overview of the post-test scores can be seen in Figure 7, with descriptive statistics available in Appendix H.

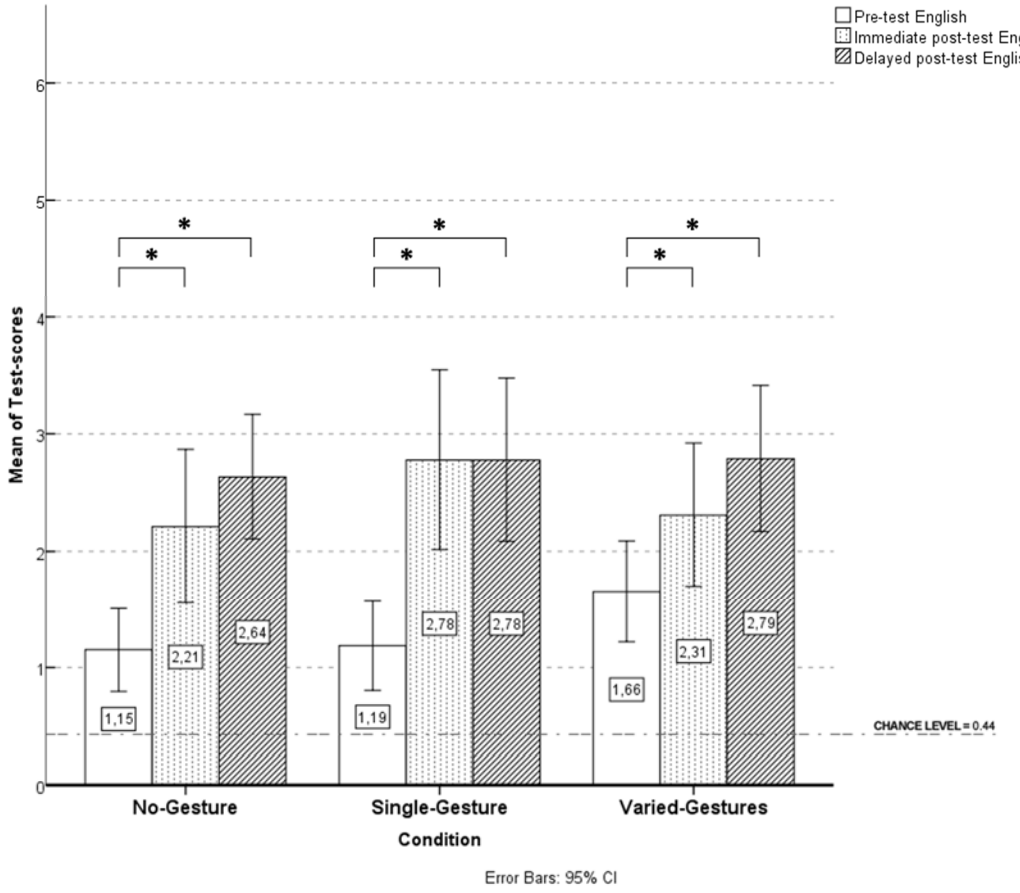


Figure 7. Bar graph showing the mean of the test-scores between the three conditions. Maximum score out of six.

* $p < .001$

4.2.2 Engagement in Robot Interaction

For all values in this section, negative numbers denote a negative intensity value of the construct (i.e. less engaged, fewer positive emotions, etc.), while positive numbers denote a positive intensity value of the construct. Furthermore, this section will look at both the difference in the averaged engagement score over the duration of the training (round 4 + round 24) as well as the difference in engagement between the two rounds separately (i.e., the decline-rate). A complete overview of all descriptive statistics can be found in Table 5.

Average Engagement

To look at the children's average engagement between the three conditions (Figure 8), a one-way MANOVA was performed, with condition as independent factor and average engagement score per construct as dependent measures. No multicollinearity was observed between the four constructs, with a maximum r =

0.595. The MANOVA found a significant effect for condition, *Wilk's* Λ = .57, $F(8, 176) = 7.16$, $p < .001$, $\eta_p^2 = .245$. As can be seen in Table 3, a significant effect was only found for child-robot engagement, when considering the adjusted $p < .0125$ significance level for multiple hypothesis testing.

Table 3. One-way MANOVA Between-Subjects effects

Factors	*df*	*F*	*p*	η_p^2
Child-Task Engagement	2, 91	1.19	.159	.040
Child-Robot Engagement	2, 91	18.87	< .001	.363
Arousal	2, 91	1.16	.015	.089
Valence	2, 91	0.52	.055	.062

Note. To adjust for Type I errors arising due to multiple hypothesis testing, a $p < .0125$ is used as significance level.

A Post-Hoc Bonferroni analysis revealed that child-robot engagement was significant between the no-gesture and single-gesture conditions, $M_{dif} = 0.91$, $p < .001$, as well as between the no-gesture and varied-gestures conditions, $M_{dif} = 0.97$, $p < .001$. No difference was found between the single- and varied-gestures conditions, $M_{dif} = 0.06$, $p = 1.000$. A complete overview of all descriptive statistics can be found in Table 5.

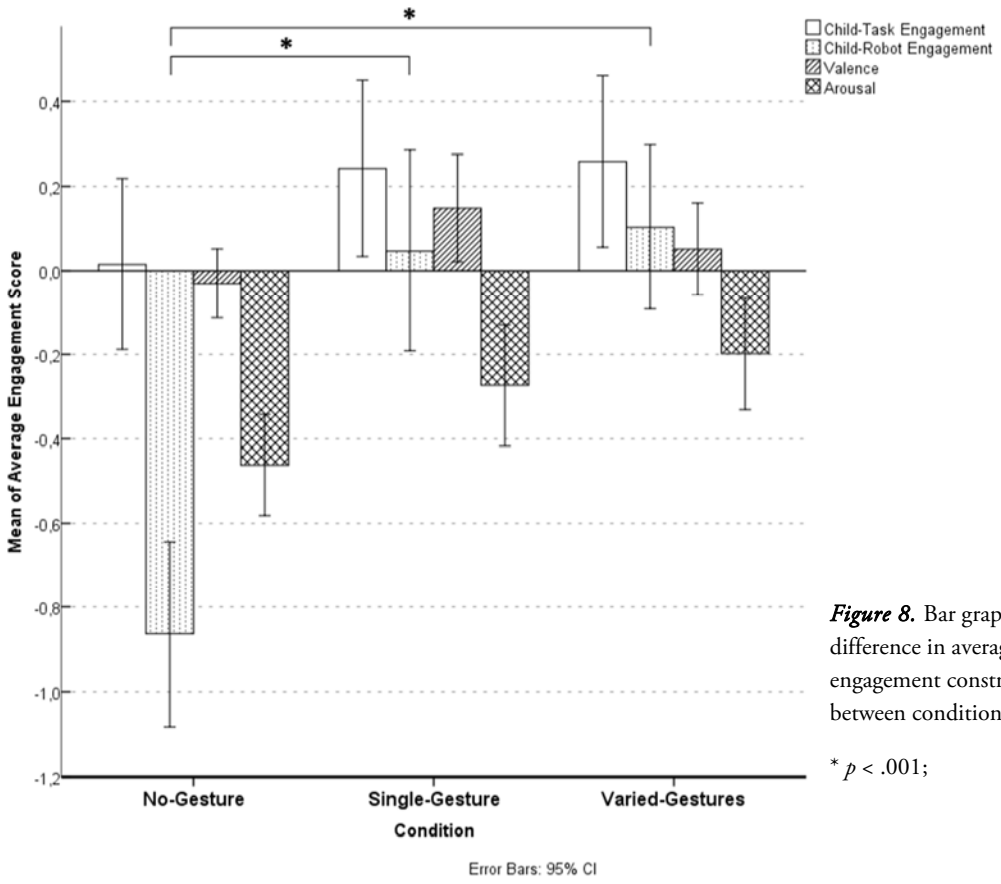


Figure 8. Bar graph showing difference in averages of engagement constructs between conditions.

* $p < .001$;

Engagement decline, between rounds

To examine the difference in engagement between the two rounds, a one-way repeated-measures MANOVA was performed, with condition as independent factor and scores for each construct in round 4 and round 24 as within factors (referred to as time). Again, to adjust for multiple hypothesis testing, a $p < 0.0125$ is required. The MANOVA revealed a significant within-subjects effect for time, *Wilk's* Λ = .19, $F(4, 88)$ = 95.56, $p < .001$, η_p^2 = .813. As can be seen in Table 4, effects were found for all constructs, indicating that they all showed significant decline in engagement between rounds (Table 5). No interaction effect was found between time and condition, *Wilk's* Λ = .875, $F(8, 176) = 1.52$, p = .153, visualised in Figure 9.

Table 4. One-way Repeated-Measures MANOVA Within-Subjects effects

Factors	*df*	*F*	*p*	η_p^2
Child-Task Engagement	1, 91	132.26	< .001	.592
Child-Robot Engagement	1, 91	134.79	< .001	.597
Arousal	1, 91	219.15	< .001	.707
Valence	1, 91	19.06	< .001	.173

Note. To adjust for Type I errors arising due to multiple hypothesis testing, a $p < .0125$ is used as significance level.

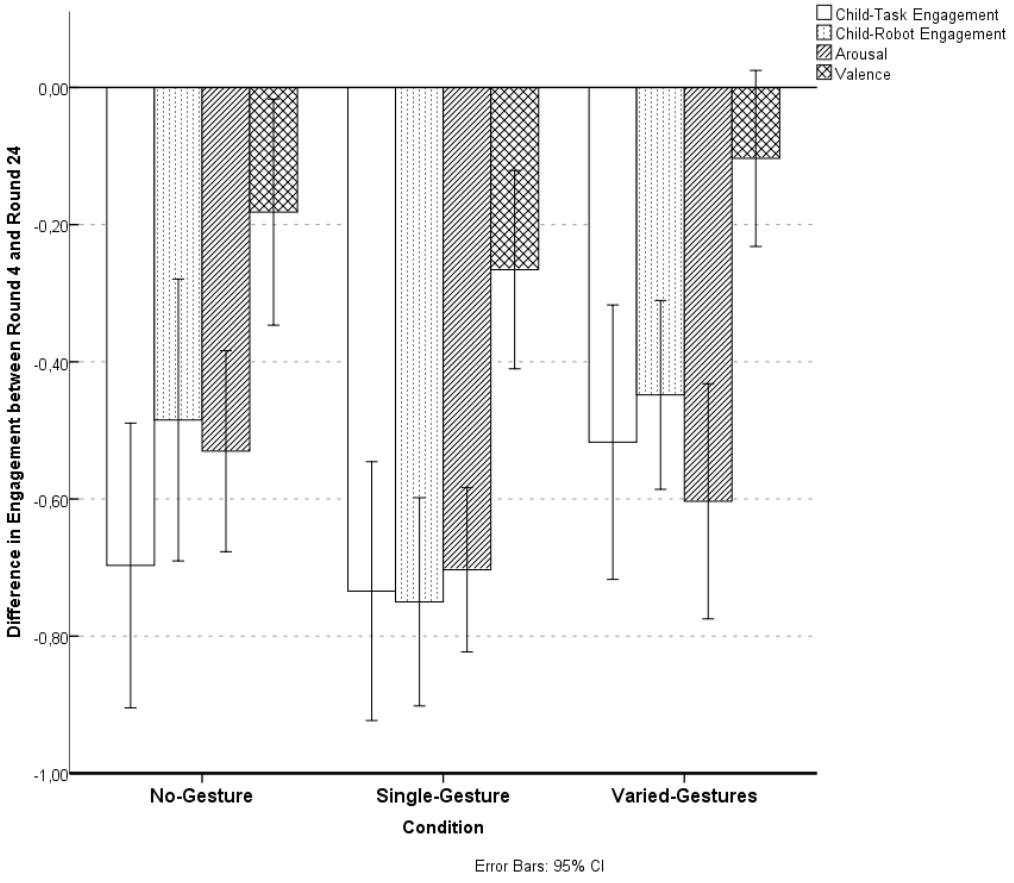


Figure 9. Bar graph showing difference scores of engagement constructs between rounds, per condition.

Table 5. Engagement scores per condition, *M (SD)*

Factors	*No-Gesture*			*Single-Gesture*			*Varied-Gestures*		
	Round 4	Round 24	**Average**	Round 4	Round 24	**Average**	Round 4	Round 24	**Average**
Child-Task	0.36 (0.58)	-0.33 (0.70)	**0.02 (0.57)**	0.61 (0.53)	-0.13 (0.72)	**0.24 (0.58)**	0.52 (0.49)	0.00 (0.68)	**0.26 (0.53)**
Child-Robot	-0.62 (0.76)	-1.11 (0.60)	**-0.86^{AB} (0.62)**	0.42 (0.64)	-0.33 (0.75)	**0.05^{A} (0.66)**	0.33 (0.47)	-0.12 (0.61)	**0.10^{B} (0.51)**
Arousal	-0.20 (0.39)	-0.73 (0.40)	**-0.46 (0.34)**	0.08 (0.46)	-0.63 (0.40)	**-0.27 (0.40)**	0.10 (0.47)	-0.50 (0.35)	**-0.20 (0.35)**
Valence	0.06 (0.37)	-0.12 (0.28)	**-0.03 (0.23)**	0.28 (0.36)	0.02 (0.45)	**0.15 (0.35)**	0.10 (0.41)	0.00 (0.23)	**0.05 (0.29)**

Notes. For each construct, scales range from -2 to +2. Positive numbers denote a positive engagement, vice versa for negative numbers. All scores between rounds were significant within a particular condition; letters denote significant differences between conditions.

4.3 Secondary Pre-registered Anthropomorphism Results

To explore whether perceived anthropomorphism changed in children due to the robot displaying gestures, a repeated-measures ANOVA was performed. The ANOVA showed that children did not differ significantly between the pre-test (M = 7.32, SD = 2.49) and the post-test (M = 7.32, SD = 2.71), $F(1, 91) = 0.01$, p = .976. When taking conditions into consideration, the varied-gestures condition had a slightly higher average than the no-gesture condition and the single-gesture condition (Table 6), however, this difference was not significant, $F(2, 91) = 0.16$, p = .856. In essence, gestures did not contribute to an overall difference in the anthropomorphisation of the robot by children. Finally, a one-way repeated-measures MANOVA showed that children also did not differ significantly between pre- and post-tests on the biological traits ($F(1, 91)$ = 0.25, p = .621), nor on the mental traits ($F(1, 91) = 0.31$, p = .579). No interaction effects were found between conditions, and testing time: Biological traits, $F(2, 91) = .02$, p = .984; Mental traits, $F(2, 91)$ = 0.30, p = .743.

Finally, a more elaborate analysis of some of the descriptive results can be found in Appendix I. These can mostly be summarised as follows: overall, anthropomorphism was largely stable between pre- and post-tests, however, large variances were found between children individually. Furthermore, children generally ascribed more mental components than biological components to the robot.

Table 6. Descriptive statistics of Perceived Anthropomorphism per condition, *M (SD)*

Factors	*No-Gestures*		*Single-Gesture*		*Varied-Gestures*	
	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test
*Overall Anthropomorphism**	7.30 (2.81)	7.21 (2.99)	7.34 (2.15)	7.28 (2.30)	7.31 (2.54)	7.48 (2.85)
*Biological Traits ***	2.82 (1.67)	2.73 (1.68)	2.50 (1.50)	2.44 (1.52)	2.69 (1.51)	2.66 (1.78)
*Mental Traits ****	4.48 (1.62)	4.48 (1.66)	4.84 (1.08)	4.84 (1.08)	4.62 (1.32)	4.83 (1.44

Notes. * Score out of 11, "Other" questions are excluded; ** Score out of 5; *** Score out of 6

4.4 Exploratory results

Whilst exploring the dataset and crosschecking findings with the experiment observation notes, an interesting discovery was made. As can be seen in Figure 10, a positive trendline is visible where children who are older tended to score higher on the immediate post-test than younger children. A one-way ANOVA with age (age 4, 5 or 6) as a factor supported this discovery, $F(1, 93) = 5.255$, $p = .007$, $\eta^2 = .103$.

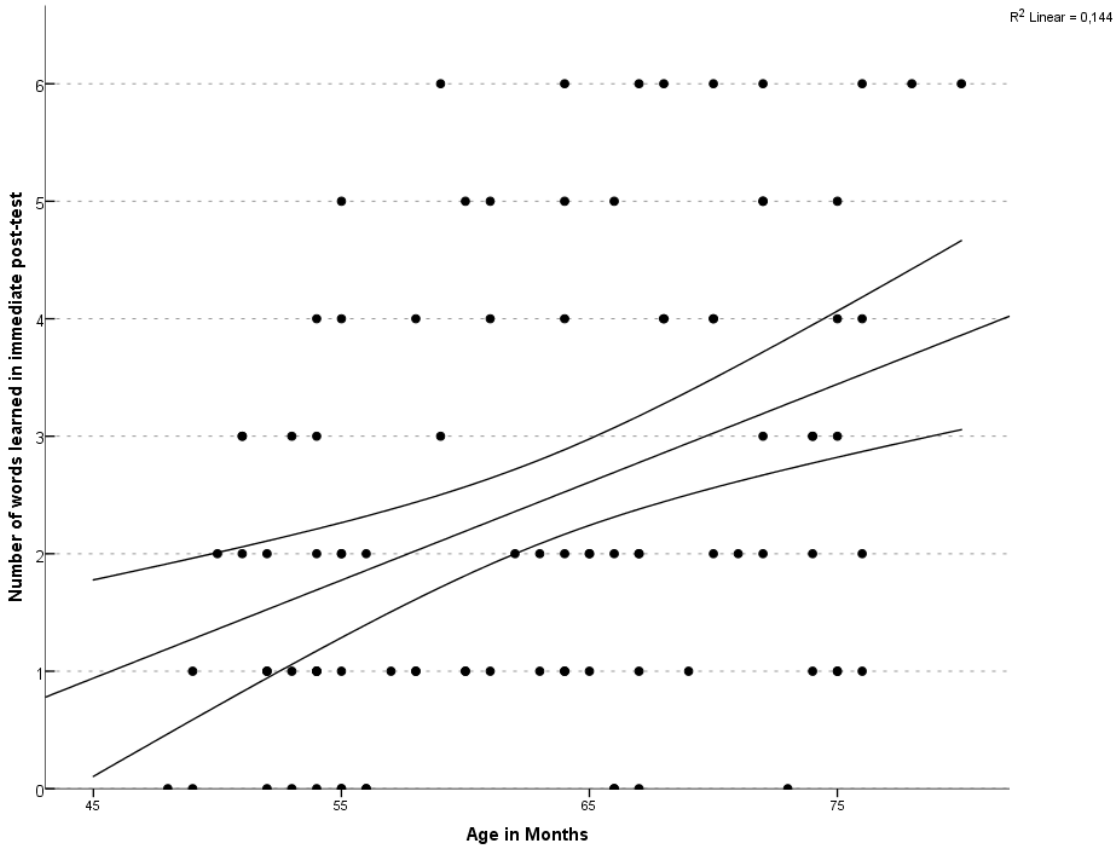


Figure 10. Scatterplot showing number of words learned per child, when looking at age.

A Post-Hoc Bonferroni analysis revealed a significant effect between the four year olds and the six year olds (M_{dif} = 1.56, p = .007). No effect was found between four-year olds and five-year olds (M_{dif} = 0.82, p = 0.126) or between the five-year olds and six-year olds (M_{dif} = 0.74, p = .307). The current section will explore whether these differences are present in other factors as well. In order to do that, the group of participants will be further divided into two smaller subgroups: four-year olds and five- and six-year olds. The five- and six-year olds are placed together in order to maintain a more consistent group distribution (as the six-year old group itself is relatively small) as well as showing a smaller difference in means. An overview of the new group distributions is visible in Table 7.

Table 7. Total participant distributions per condition and age group

Factors	*No-Gestures*	*Single-Gesture*	*Varied-Gestures*	**Total**
Age 4				
n	14	14	10	**38**
% girls	50.0	42.9	70.0	**52.6**
% boys	50.0	57.1	30.0	**47.4**
Age in Months				
M (SD)	4 years, 6 months (*SD* = 5 months)	4 years, 6 months (*SD* = 2 months)	4 years, 6 months (*SD* = 2 months)	**4 years, 6 months (*SD* = 3 months)**
Ages 5 and 6				
n	19	18	19	**56**
% girls	47.4	44.4	52.6	**48.2**
% boys	52.6	55.6	47.4	**51.8**
Age in Months				
M	5 years, 9 months (*SD* = 5 months)	5 years, 8 months (*SD* = 7 months)	5 years, 9 months (*SD* = 5 months)	**5 years, 9 months (*SD* = 5 months)**

4.4.1 Word Learning per Age Group

To test for the influence of age on the word learning scores, a one-way repeated-measures ANOVA was performed between conditions and age groups (independent factors) and the three testing scores (time; dependent factors).

A significant interaction effect was found for time and age groups, $F(2, 176) = 5.80$, p = .004, η_p^2 = .062. As can be seen in Table 8, children aged five and six scored performed better on word learning than children aged four. Means also show a trend where children aged five- and six showed higher scores in the gesture conditions than younger children (Figure 11), however, no overall significant interaction effect was found between conditions, age groups and testing time, $F(4, 176) = 2.41$, p = .051.

Table 8. Vocabulary learning scores per condition and age group, *M (SD)*

Factors	*No-Gestures*	*Single-Gesture*	*Varied-Gestures*	**Total**
Pre-test				
Age 4	1.14 (0.95)	0.71 (0.61)	1.60 (1.08)	**1.11 (0.92)**
Ages 5 and 6	1.16 (1.07)	1.56 (1.20)	1.68 (1.20)	**1.46 (1.16)**
Immediate Post-test				
Age 4	2.14 (1.99)	1.29 (0.99)	2.00 (1.70)	**1.79 (1.61)**
Ages 5 and 6	2.26 (1.79)	3.94 (2.04)	2.47 (1.58)	**2.88 (1.93)**
Delayed Post-test				
Age 4	2.57 (1.65)	1.50 (1.29)	1.80 (1.03)	**1.97 (1.42)**
Ages 5 and 6	2.68 (1.42)	3.78 (1.77)	3.32 (1.67)	**3.25 (1.65)**

Note. Chance level was calculated at 0.44 words.

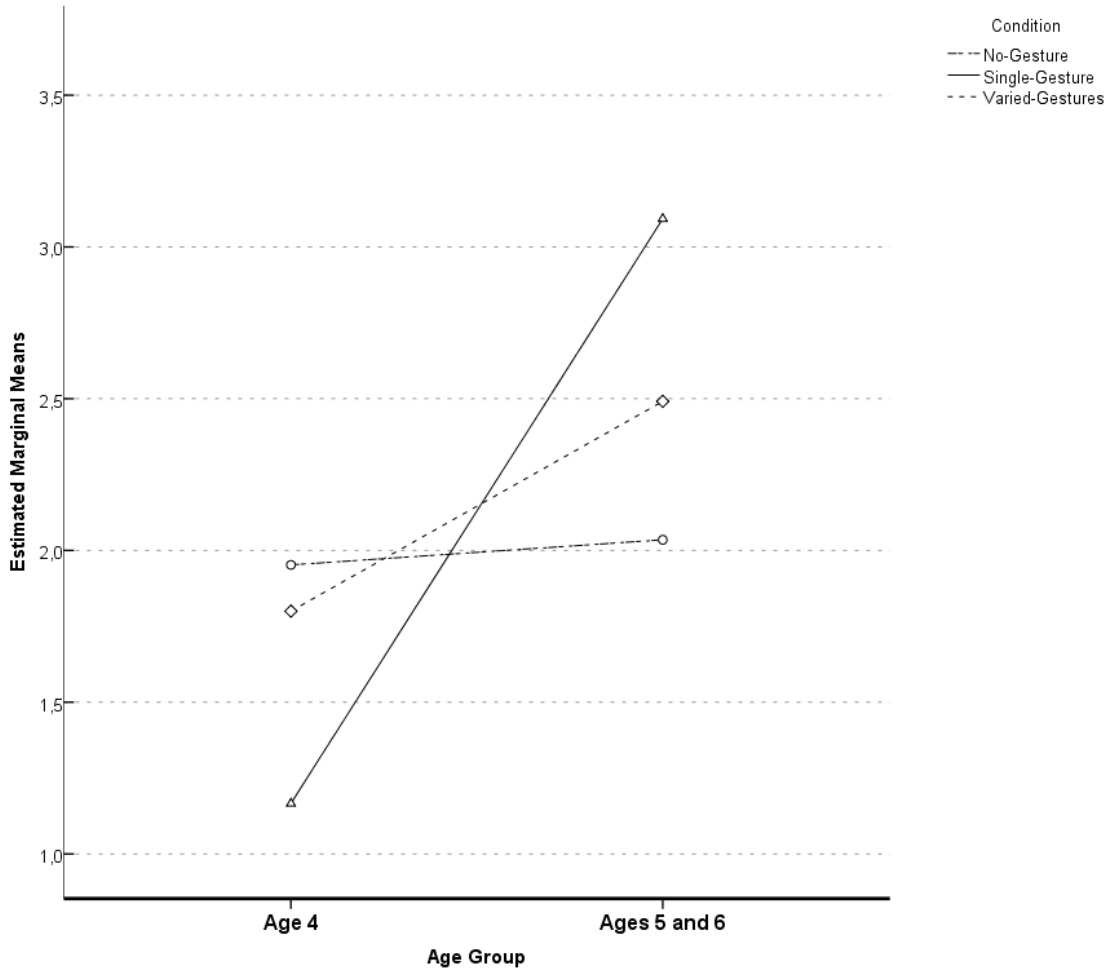


Figure 11. Estimated Marginal Means of pre-test, immediate post-test and delayed post-test between age groups, per condition

4.4.2 Engagement per Age Group

As the results for engagement per age group are fairly elaborate, only the main points will be spoken about here. For a more complete overview of the results, please refer to Appendix J.

Overall, no main effects for age were found for any of the four constructs (child-task engagement, child-robot engagement, arousal, and valence). Similarly, no interaction effects were found between age group and condition, apart from valence, $F(2, 88) = 5.93$, $p = .004$, $\eta_p^2 = .119$. A simple effects analysis indicated that children aged five and six showed more negative emotions ($M = -0.07$, $SD = 0.20$) in the varied-gestures condition than children aged four ($M = 0.28$, $SD = 0.30$), $M_{dif} = 0.34$, $p = .003$. For engagement decline, similar results were found as in the main analysis; main effects were found for all constructs, indicating that for both age groups engagement declined. However, no interaction effects were found, indicating that the two age groups did not differ significantly from each other.

4.4.3 Perceived Anthropomorphism per Age Group

Again, an overview of these results can be found in Appendix K. In general, however, these results were identical to the secondary analysis; a repeated-measures ANOVA for both four-year olds and five- and six-year olds showed no significant differences in anthropomorphism, nor were there any significant interaction effects between age and conditions.

4.4.4 Completion Time and Error-Rate

A more elaborate overview of these results can be found in Appendix L. Mostly, these can be summarised as follows: Children in the gesture conditions ($M = 17.80$ minutes, $SD = 4.48$ minutes) saw significantly longer experiment durations compared to the no-gesture condition ($M = 14.31$ minutes, $SD = 2.43$ minutes). Children aged five and six also saw a significantly shorter duration in the single-gesture condition than those aged four ($M_{dif} = 5.52$ minutes, $p < .001$).

In terms of error-rate, children aged five and six ($M = 9.64$ errors, $SD = 6.90$) made fewer errors than children aged four ($M = 15.47$ errors, $SD = 7.89$). A trend was visible that the single-gesture condition made fewer errors than the other conditions, which was significant for the five- and six-year olds.

5. Discussion

In this study an experiment was conducted to explore the effects of gestures, repetitive or varied, on children's word learning performance and retention, their engagement with a robot, and finally their perceived anthropomorphism of a robot. It was hypothesised that gestures in general would lead to (H1a) higher word learning performance, (H1b) better retention and (H3) increased engagement with the robot. Furthermore, it was expected that by adding variation to the robot's gestural production, (H2a) word learning performance and (H2b) retention, as well as (H4) engagement, would increase even further compared to repeated gestures.

5.1 Word learning

In general, all children that participated in the study, regardless of condition, learned new English words from their interaction with the robot. Based on the previous study by de Wit et al. (2018), as well as the dual coding theory (J. M. Clark & Paivio, 1991) and the cognitive development theory (Piaget, 1952), it was expected that gestures would have a positive effect on the ability of children to learn new words. However, no such effect was found in the current study, and neither H1a nor H1b was supported. What is interesting, is that this contrasts with de Wit et al. (2018), where gestures did see a difference. The explanation for this could be twofold: on the one hand, the updated testing measures (using three rounds of six words in the word learning tests, instead of one round of six words) may have improved robustness and caught more false positives (in the form of guessing). On the other hand, the present study used slightly different stimuli in the form of objects (bridge, stairs, pencil, and spoon), which may have been more difficult words to grasp compared to the all-animal design in de Wit et al. (2018). In terms of variations in gesture usage, the current study is inconclusive. While means did show a positive trend in terms of word retention, showing higher scores for recall than the immediate post-test, these results were not significant. Thus, no support was found for H2a and H2b.

Interestingly, as the later exploratory results showed, age seemed to play a substantial role on word learning performance. Between conditions, no significant differences could be found. But as can be seen in Figure 11, a trend was visible where children aged five or six saw a substantial benefit from the single-gesture condition compared to their peers in the other conditions. This particular subgroup also showed substantially lower training durations, partially caused by a lower error-rate. Taken together, this is interesting, as it does imply a trend that gestures can help in learning with robot tutors. Curiously, this trend was not visible as strongly in the varied-gestures condition. However, this may be somewhat linked to the design of this study. Several studies have argued that repetition is beneficial for second language learning (e.g., Ghazi-Saidi & Ansaldo, 2017; Lambert, Kormos, & Minn, 2017). In this case, the innate repetition

of the single-gesture condition may have given children in this condition an advantage, while seeing a new gesture each time in the varied-gestures condition may have inhibited this repetition process. However, that does not mean that variations in gestures could not still be helpful, but perhaps this too needs to be combined with repetition. For a more conclusive result, a longitudinal study would perhaps be more appropriate.

In general, the means seem to suggest that there is a difference in learning capabilities between four-year olds, and five- and six-year olds. However, the no-gesture conditions for both age group showed nearly identical results, with larger, opposite, differences being visible in the gesture conditions. In both gesture conditions, four-year olds scored lower than their no-gesture peers, while the opposite was visible in the five- and six-year old group. Again, this provides an indication that there is a difference in how gestures are interpreted between the two age groups. For some reason, older children showed a considerably larger learning benefit when presented with gestures. A study by Novack, Goldin-Meadow and Woodward (2015) may provide further insight; in their study, they taught two- and three-year old children to perform an action on a toy via iconic gestures. Both age groups showed that they could learn from gestures, but the three-year old group showed a much larger interpretation effect of the gestures and subsequently higher performance. The discrepancy that is visible between age groups in the current study may be an extension of what had happened in Novack et al. Similar to their study, younger children may have had more difficulties with the representative aspects of the gestures, finding it difficult to combine it with their own knowledge or perhaps caused by their lack of knowledge.

This does however pose an interesting discussion regarding the cognitive development theory (Piaget, 1952) and the dual-coding theory (J. M. Clark & Paivio, 1991). In a sense, the current outcome is predicted by Piaget's theory, as older children likely have a larger pool of knowledge to refer back to, based on their vaster life experiences. Therefore, gestures may give them a better chance at giving the correct answer. However, the dual-coding theory states that humans should learn better when being presented with more modalities, regardless of their experience. For younger children at least, the present study seems to indicate an opposite effect when presented with gestures.

Partially, this may be explained by the *Cognitive Load Theory* (Sweller, 1988; Sweller, Ayres, & Kalyuga, 2011), which states that all humans have a limited capacity for cognitive processes, such as working memory capacity. Typically, dual-coding has been found to increase learning, as it can assist with offloading cognitive processes to different parts of the brain that are linked to visual and verbal coding, leading to lower cognitive load (Mayer & Moreno, 2003). However, there have been cases where applying dual-coding to language learning for children when using images, actually led to lower word learning scores (e.g., Acha, 2009). Acha found that naming words based on visual representations, rather than on the words themselves,

led to higher recognition times. It may thus be possible that for young children, the attempt to recognise words from gestures puts additional strain on their cognitive capacities, leading to lower performance.

Furthermore, when schemas have not yet had the chance to fully form, it costs a larger amount of cognitive ability to process new information, potentially reaching the maximum processing capability (Sweller, 1994). Yet this maximum capacity is not set in stone, it grows as children age (Cowan, 2016). And while previous studies have shown that gesturing can decrease cognitive load (e.g., Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001; Ping & Goldin-Meadow, 2010), these have mostly focused on older children (around age eight to ten). As seen in the present study, the capabilities of very young children differ widely compared to their older peers. It may very well be possible that the combination of a robot, gestures and the very young age of the children have caused an overload in their cognitive capabilities. Interestingly, while existing research seems to allude to cognitive load being a possible cause of lower learning scores in social robot studies, there seem to be very few studies that actually focus on the effects of social robots on cognitive load; a definite point of attention for future studies, even more so for studies where the robot is very salient in the interaction.

Observations showed further evidence of young children's difficulties with gesture interpretation. During the initial introduction of the game, the robot explained the game by first asking for the word "paard" (Dutch for "horse"). The children were then presented with a selection of three images. In almost all cases, the children correctly identified the image of the horse. The robot then immediately followed up with the English practice round, this time asking for a "horse". A large portion of the children provided a wrong answer here. For the non-gesture condition, this is to be expected as they had not yet learned the word. However, in the gesture conditions, the robot accompanied both the Dutch and English rounds with an identical horse-riding gesture. Despite having previously seen the exact same gesture in the Dutch round, accompanied by the "paard" word they knew, the gesture did not permit them to link the new English word to the same gesture. Thus, children may view the gestures as being a separate entity from the English words. Future variations of this gesture study may wish to more actively incorporate the gestures in the learning experience, by asking the children to perform them together with the robot, perhaps during the concept mapping phase. As previously mentioned, several studies have shown that gesture (re-)enactment can improve learning scores (Macedonia & Knösche, 2011; Tellier, 2008). Again though, this phenomenon is subject to individual differences, as a select few children did see benefits from gestures. These children recognised the horse-riding gesture that the robot portrayed and attempted to press the correct image even before the robot uttered the word "paard" or "horse". Some children even said "Hey, I have seen this before, that's a horse!" before promptly pressing the correct image. In later conversations with the children, some

of them noted that they joined a horse-riding club or had a particular interest in horses, showing that gestures do allow children to refer back to their existing knowledge.

In conclusion, while the overall performance scores showed that children had no benefit from gestures, a more specific look at differences in ages told a different story; older children seem to have a benefit from gestures compared to younger children. While there are possible explanations on why this happened, the current study does not have the ability to confirm these explanations. What is an important take-away is that at these very young ages, a large gap can be found between children differing in age as little as a year. It may very well be possible that an entirely different approach is needed for these children, to further optimise the effectiveness of robots in education.

5.2 Engagement

5.2.1 Measurements of Engagement

For the second set of hypotheses regarding engagement, partial support for the third hypothesis could be found. This difference was mostly visible in child-robot engagement, which saw fairly large differences between conditions. Children in either of the gesture conditions saw significantly higher average engagement with the robot than those in the no-gesture condition. Similar trends could be seen for valence and arousal, though these were not significant. Overall, this does seem to suggest that a robot's motion can lead to a longer sustained engagement of the child. Interestingly, no effect was found for child-task engagement, but perhaps this is also to be expected. Mostly, the task itself took place on the tablet and was consistent between the conditions. Whenever the child was asked to perform an action on the tablet, the robot demanded no further attention from them; the robot said nothing and made no salient movements. In other words, the perception of the task was perhaps identical for all children regardless of which condition they were in.

In terms of engagement decline, every construct saw a drop. Curiously, no differences were found between the three conditions. The data, however, seems to suggest that rate at which engagement declined was greatest for the single-gesture condition, while the no- and varied-gesture conditions saw similar declines. It should be noted that training duration may have been an additional variable at play here. On average, both gesture conditions took several minutes longer than the no-gesture condition. It stands to reason that if engagement drops over time, it has the opportunity to drop even further when the training takes longer, assuming no ceiling has been met. On the other hand, as both gesture conditions saw similar durations, this does seem to imply that a variation in gestures can keep a child's engagement for longer, as alluded to in Tanaka et al. (2007) and Ros et al. (2014). The current study is not capable of further answering this notion, but future variations of this study may wish to look into controlling for duration when measuring engagement.

In terms of age, no effects were found between the two groups, indicating that engagement was similar for all children regardless of their age. Interestingly, an interaction effect was found for valence between the two age groups in the varied-gestures condition. However, taking in regard the entire scale (-2 to +2), this difference was fairly moderate. Finally, all age groups saw similar declines, and no effects were found between condition and age group.

5.2.2 Observation of Engagement

Overall, many children seemed to enter the training fairly positively, happily reacting to seeing Robin again and sometimes even trying to engage into conversation with the robot. Seen from anecdotal observations, most children answered the very first question asked by Robin ("Do you still know me?") with a sound "yes!". During the interaction, some children asked the robot questions (both task-specific ("Could you repeat that word?", "Was that a horse, Robin?") and non-task specific questions ("Can you play soccer, Robin?")), yet these questions were never reciprocated by the robot. As they slowly learned that the robot would continue playing the game, never reacting towards something they say or do, their attention slowly wavered. In part, this may be explained by the somewhat unnatural interaction the children had with the robot. Anecdotally, the children seemingly expected a form of interaction with the robot, a 'dialogue' instead of the 'monologue'. Related results can be seen from the anthropomorphism questionnaire, with many children noting that they would like to 'do' something together with Robin (e.g. dancing, playing football).

It is likely that this lack of (re)active response is what currently sets the robot apart from a human teacher. Future research may wish to look into more active responses from the robot. Even something as 'simple' as using face recognition that recognises when a child looks away with the robot responding "Hey, keep your attention" may already let the children feel that the robot is a more living, responding entity. It may very well be that the highly social nature of the robot in the group introduction set a baseline of expectations for the children, expectations the robot could not meet on a technological level**.** Children then lowered their expectations quickly as soon as they found out that their interactions towards the robot were not reciprocated, lowering their emotional and activity output to a fairly neutral state.

In many ways, this is in line with what was previously stated in Duffy (2003) and Bartneck et al. (2004); children had certain expectations that were never met by the robot. To a certain extent, this may have actually been caused by the design of the group introduction. In an attempt to alleviate anxiety of the robot, the knife may have cut on both sides by also bringing about certain expectations in the form of the interaction. During the group introduction, the interaction was very much going two ways. The robot would ask "Do you want me to dance?" before the children would expressively shout "yes!". The robot then asked, "Will you join me then?" eliciting an identical response from them. This may have caused an illusion of choice and active involvement for the children, which never returned in the training itself.

5.3 Anthropomorphism

Interestingly, anthropomorphism saw no differences across the board, regardless of condition or age. Therefore, there is no clear answer to what extent gestures influence anthropomorphism. Children tended to show a fairly high degree of anthropomorphism in many cases, resulting in a final score of 7.32 out of eleven in both pre- and post-tests. On a whole, it may be possible that this is caused by the so-called novelty effect. This effect is a widespread 'issue' in all short-term robot studies; as children meet a robot for the very first time, they can become mesmerised and show increased interest in it. Especially for such a young group of children, having a robot come to school seemed to be a very exciting prospect. This excitement may have overloaded them with positive feelings regarding the robot, skewing the results in the current study. This is further substantiated by the fact that regardless of whether they liked the training or not, all children seemed to want to play with the robot again, though not necessarily with the current game they had played.

The current study saw near identical patterns as in van den Berghe, de Haas et al. (2019). Here too, large individual variances between children were found, as explained earlier by Epley et al. (2008). Similar to the study by van den Berghe, de Haas et al., average anthropomorphism in both pre- and post-tests saw stable scores. What is even more intriguing is how children attributed more mental than biological traits to the robot (again, similar to van den Berghe, de Haas et al. (2019)). In a sense, this is to be expected. Children often noted a fairly mechanical viewpoint in regard to the biological part of the questionnaire. Robin is made of metal, therefore he cannot grow. He has no (or a very small) mouth, therefore he does not need food. While with the mental component, children seemed to relate the robot more directly to their own referential world. He can be sad, if he falls while playing. He is smart, as he went to school too. This is also in line with previous research that suggest that children tend to more easily assign emotional characteristics to robots, rather than cognitive and behavioural characteristics (Beran et al., 2011).

This similarity between the current study and van den Berghe, de Haas et al. (2019) is however interesting. In many ways, it speaks against the previous assumption of the novelty effect occurring. It shows that anthropomorphism, as it is currently measured, does not change depending on whether it is a longitudinal study (like van den Berghe, de Haas et al. is) or a single-session study. Partially, this may be explained due to the nature of the questions asked. Questions like "Do you think Robin can see things?" seem to have a fixed answer based on the perception of eyes. In both pre- and post-tests, results seemed to support this notion, as nearly all of them answered "Yes, he has eyes" here. This does however beg the question whether children link their expectations directly to physical attributes as mentioned in Bartneck et al. (2004). The current study is however not equipped to accurately assess this, as none of the questions directly related to things that happened in any of the conditions. A future study may wish to see if anthropomorphism changes if a robot actively defies expectations for children. For example, by pre-testing

to see if children perceive the ability of robots to see things and then making an obvious case for the fact that it cannot see. If post-test anthropomorphism changes, then this would provide further evidence that the currently used questionnaire is appropriate for measuring anthropomorphism for young children.

It should further be noted that there are other limitations to the current measurements of anthropomorphism. Naturally, as was discussed earlier in this study, measuring anthropomorphism in young children is difficult and provides several reliability challenges. A binary scale provides little leeway in terms of accuracy of a child's actual response, yet at the same time using standardised Likert scale questionnaires is also an impossibility for these young children. This is further exacerbated by the fact that the questions used were fairly concrete, even though they try to represent an abstract construct that children would not be able to understand. As of yet, it is unknown whether the current questionnaire and its setup are appropriate to accurately measure anthropomorphism.

Furthermore, the present questionnaire was limited in scope. It contained no questions that attempted to understand the child's perception of communicative abilities in the robot. It may very well be possible that a robot using gestures may be perceived as being more capable of communication. However, that again begs the question of abstract constructs: how does one ask a four-year old child on how well a robot can communicate?

Finally, a more practical limitation is found in the way the post-test questionnaire was held. During this questionnaire, both the experimenter as well as the robot were present. If the children truly perceive the robot as a social agent, this may have resulted in the children giving socially desirable answers. In general, the consistency between pre- and post-tests would argue that this may not have happened, as the robot was not present in the pre-test. Nevertheless, future studies may wish to look into the effects of anthropomorphism where the robot is either present or not, to gain more insights in whether children see the robot as a social, human-like agent that could potentially be 'hurt' by negatively framed answers.

5.4 General Limitations

While care was taken to prevent issues during the design of this study, there are however still a few limitations. First of all, while the design had hoped to keep the training as short as possible, by placing the group introduction and pre-tests on different days from the experiment, some of the children had substantially long experiment durations. In a few cases, the experiment from the child entering the room to leaving it again, took over 45 minutes. Training fatigue due to the long durations was very apparent: children got extremely bored and seemed to not always take the post-tests seriously. Some answered the anthropomorphism questionnaire with "I don't know" on many questions despite having answered them more seriously in the pre-test. Others seemed to just randomly tap the screen during the post word learning

test, or even just press the same image in a single corner for each round. Future studies may wish to further prevent this long duration by taking multiple sessions or keeping the experiment from growing stale by adding in additional, non-task specific interactions with the robot.

Similarly, while a deliberate choice was made to have the experiment take place on schools, this may have brought about additional distractions for the children. Phenomena such as break times taking place during the experiment, children outside of the experiment room trying to gain attention of the robot or the child being trained, teachers entering rooms or being present during the experiment, all added additional distractors for the children. While in many ways this provides ecological validity by placing the robot in an active school environment, it is unknown what kind of effects these distractors may have had on the results. Furthermore, as there was a delay between each testing point in time, it is unknown what kind of effect this may have had. Similar to Vogt et al. (2019), children went home and, in many cases, spoke to their parents about their experiences with the robot. There is a distinct possibility that the children may have spoken about what words they had learned, inadvertently repeating these words more often in their head leading to higher post-test scores. Similarly, the current testing measurements do not provide insights in whether the children learned from the pre-test itself.

Care should also be taken in the generalisation of these results. All participants in the current study were natively Dutch. Previous research has shown that one's native language can have a large effect on how well one can learn a particular second language (Shatz, 2016). In other words, teaching native Dutch children English can lead to very different results compared to teaching Japanese children English. Similar results have been found for robot perception, where constructs such as likeability, engagement and trustworthiness of the robot are vastly dependent on the culture of the participant (Bartneck, Suzuki, Kanda, & Nomura, 2007; Li, Rau, & Li, 2010).

Finally, it was also readily apparent during all of the trainings that each child differed enormously from another. Some can handle a lot of background noises; others need a quieter working area. Some can handle repetitive tasks; others need more stimulation. In other words, the current one-size-fits-all approach of the robot interaction is perhaps not suitable to accommodate the different approaches some children need. The present setup of the training disregards something a teacher can give: a personalised approach based on the child's needs. Background information of each child, such as personality, and individual skills and abilities, is currently discarded. Yet this information can have a sizeable effect on how well the children learn, for example, lower language skills have been found to change the effectiveness of gestures in language learning (Post, Van Gog, Paas, & Zwaan, 2013). It would likely benefit the robot-interaction as a whole if the system could provide a more active incorporation of this already present information. Future research may benefit from categorising children (establishing this category together with their teacher, e.g. a shy child,

a restless child, etc., but also in regard to their current skill levels and how well they learn) and then giving them an approach more specifically tailored to their needs.

5.5 Conclusion

This study set out with two research questions: how a robot's gesture use would affect second language word learning, engagement, and perceived anthropomorphism in children, and how a variation in gesture production would affect these constructs. Therefore, a robot in a classroom-like setting was used to teach children six English words through either using no gestures, a single gesture that was repeated for each unique word, or a new gesture variation for every subsequent time a word was presented.

Based on these results, it was concluded that while children generally learned from the interaction with the robot, gestures (and their variations) initially did not seem to increase word learning capabilities. However, a more thorough analysis revealed that there was a large discrepancy in word learning between ages. While gestures did not specifically see significant results, the data does show trends where gestures seem to help older children learn more effectively, though no such differences were visible for varied gestures. Engagement did see differences in favour of both gesture conditions, and it too saw discrepancies between ages. Here, gesture variations saw a few significant results, though overall these were fairly minor. Finally, anthropomorphism saw no differences between pre- and post-test, nor between conditions. In general, though, children anthropomorphised the robot to a large degree, which may have been caused by either the novelty effect or an unreliable testing method.

In conclusion, there are two main take-aways for this study. First of all, at these young ages, children show very different results, even when separated by as little as a year. At the very least, gestures seem to affect engagement, and possibly word learning too dependent on the age. Subsequently, this shows that there is a need for further research that looks specifically at smaller age intervals in robot interactions. The second take-away builds upon this; between, but even within certain age groups, the children have vastly different skill levels, attention spans, personalities, and expectations. Many current robotic studies disregard these differences and the roles teachers have to play in the education of children. If the HRI community wishes to portray that robots are not intended to replace teachers, but instead act as an extension to the teacher's arsenal of learning tools, then the community should also start designing robots as such. A one-size-fits-all robot is not the way forward, and more care should be taken into designing more personalised experiences that are fed by the information that is already present, in conjunction with the teachers that already know these children.

References

Acha, J. (2009). The effectiveness of multimedia programmes in children's vocabulary learning. *British Journal of Educational Technology*, *40*(1), 23–31. https://doi.org/10.1111/j.1467-8535.2007.00800.x

Alibali, M. W., & 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*, *39*(5), 349–366. https://doi.org/10.1111/j.1467-8535.2008.00890_7.x

Bajcsy, R., Aloimonos, Y., & Tsotsos, J. K. (2018). Revisiting active perception. *Autonomous Robots*, *42*(2), 177–196. https://doi.org/10.1007/s10514-017-9615-3

Barak, M., & Zadok, Y. (2009). Robotics projects and learning concepts in science, technology and problem solving. *International Journal of Technology and Design Education*, *19*(3), 289–307. https://doi.org/10.1007/s10798-007-9043-3

Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)* (pp. 591–594). IEEE. https://doi.org/10.1109/roman.2004.1374827

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

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2007). The influence of people's culture and prior experiences with Aibo on their attitude towards robots. *AI and Society*, *21*(1), 217–230. https://doi.org/10.1007/s00146-006-0052-7

Belpaeme, T., Baxter, P. E., De Greeff, J., Kennedy, J., Read, R., Looije, R., … Coti Zelati, M. (2013). Child-robot interaction: Perspectives and challenges. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 8239 LNAI, pp. 452–459). https://doi.org/10.1007/978-3-319-02675-6_45

Belpaeme, T., Baxter, P. E., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., … Humbert, R. (2013). Multimodal Child-Robot Interaction: Building Social Bonds. *Journal of Human-Robot Interaction*, *1*(2), 33–53. https://doi.org/10.5898/jhri.1.2.belpaeme

Beran, T. N., Ramirez-Serrano, A., Kuzyk, R., Fior, M., & Nugent, S. (2011). Understanding how children understand robots: Perceived animism in childrobot interaction. *International Journal of Human Computer Studies*, *69*(7–8), 539–550. https://doi.org/10.1016/j.ijhcs.2011.04.003

Burns, R., Jeon, M., & Park, C. (2018). Robotic Motion Learning Framework to Promote Social Engagement. *Applied Sciences*, *8*(2), 241. https://doi.org/10.3390/app8020241

Butcher, C., & Goldin-Meadow, S. (2010). Gesture and the transition from one- to two-word speech:

when hand and mouth come together. In D. McNeill (Ed.), *Language and Gesture* (pp. 235–258). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511620850.015

Cenoz, J. (2003). The additive effect of bilingualism on third language acquisition: A review. *International Journal of Bilingualism*, *7*(1), 71–87. https://doi.org/10.1177/13670069030070010501

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on socially shared cognition* (pp. 127–149). https://doi.org/10.1037/10096-006

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Cook, S. W., Duffy, R. G., & Fenn, K. M. (2013). Consolidation and transfer of learning after observing hand gesture. *Child Development*, *84*(6), 1863–1871. https://doi.org/10.1111/cdev.12097

Cook, S. W., & Goldin-Meadow, S. (2006). The Role of Gesture in Learning: Do Children Use Their Hands to Change Their Minds? *Journal of Cognition and Development*, *7*(2), 211–232. https://doi.org/10.1207/s15327647jcd0702_4

Corballis, M. C. (2003). From mouth to hand: Gesture, speech, and the evolution of right-handedness. *Behavioral and Brain Sciences*, *26*(2), 199–208. https://doi.org/10.1017/S0140525X03000062

Cowan, N. (2016). Working Memory Maturation: Can We Get at the Essence of Cognitive Growth? *Perspectives on Psychological Science*, *11*(2), 239–264. https://doi.org/10.1177/1745691615621279

de Haas, M., van den Berghe, R., de Wit, J., Oudgenoeg-Paz, O., Krahmer, E. J., & Vogt, P. (2019). Child Engagement during a Long-Term Robot-Assissted Language Learning Interaction. *(Unpublished)*.

de Wit, J., Schodde, T., Willemsen, B., Bergmann, K., de Haas, M., Kopp, S., … Vogt, P. (2018). The Effect of a Robot's Gestures and Adaptive Tutoring on Children's Acquisition of Second Language Vocabularies. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18* (pp. 50–58). New York, New York, USA: ACM Press. https://doi.org/10.1145/3171221.3171277

de Wit, J., Willemsen, B., de Haas, M., Krahmer, E. J., Vogt, P., Merckens, M., … Wolfert, P. (2019). Playing Charades with a Robot: Collecting a Large Dataset of Human Gestures Through HRI. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Vol. 2019-March, pp. 634–635). IEEE. https://doi.org/10.1109/HRI.2019.8673220

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

Duffy, B. R., Rooney, C. F. B., O'Hare, G. M. P., & O'Donoghue, R. P. S. (2000). *What is a Social Robot? PhD Thesis*. University College Dublin. Retrieved from

https://www.researchgate.net/publication/228803576

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When We Need A Human: Motivational Determinants of Anthropomorphism. *Social Cognition*, *26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143

Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7621 LNAI, pp. 199–208). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34103-8_20

Ghazi-Saidi, L., & Ansaldo, A. I. (2017). Second Language Word Learning through Repetition and Imitation: Functional Networks as a Function of Learning Phase and Language Distance. *Frontiers in Human Neuroscience*, *11*, 463. https://doi.org/10.3389/fnhum.2017.00463

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining Math: Gesturing Lightens the Load. *Psychological Science*, *12*(6), 516–522. https://doi.org/10.1111/1467-9280.00395

Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., … Breazeal, C. L. (2016). Affective Personalization of a Social Robot Tutor for Children's Second Language Skills. In *Proceedings of theThirtieth AAAI Conference on Artificial Intelligence* (pp. 3951–3957).

Haelermans, C. (2017). *Digital Tools in Education. On Usage, Effects and the Role of the Teacher*. Stockholm: SNS Förlag. Retrieved from https://www.sns.se/wp-content/uploads/2017/10/digital-tools-in-education.pdf

Haring, K. S., Mougenot, C., Ono, F., & Watanabe, K. (2014). Cultural Differences in Perception and Attitude towards Robots. *International Journal of Affective Engineering*, *13*(3), 149–157. https://doi.org/10.5057/ijae.13.149

Hassenzahl, M., & Tractinsky, N. (2006). User experience - A research agenda. *Behaviour and Information Technology*, *25*(2), 91–97. https://doi.org/10.1080/01449290500330331

Hickok, G. (2009). Eight problems for the mirror neuron theory of action understanding in monkeys and humans. *Journal of Cognitive Neuroscience*, *21*(7), 1229–1243. https://doi.org/10.1162/jocn.2009.21189

Hostetter, A. B. (2011). When Do Gestures Communicate? A Meta-Analysis. *Psychological Bulletin*, *137*(2), 297–315. https://doi.org/10.1037/a0022128

Huang, C.-M., & Mutlu, B. (2016). Modeling and Evaluating Narrative Gestures for Humanlike Robots. In *Robotics: Science and Systems IX*. Robotics: Science and Systems Foundation. https://doi.org/10.15607/rss.2013.ix.026

Huizenga, J., Admiraal, W., Akkerman, S., & Ten Dam, G. (2009). Mobile game-based learning in

secondary education: engagement, motivation and learning in a mobile city game: Original article. *Journal of Computer Assisted Learning*, *25*(4), 332–344. https://doi.org/10.1111/j.1365-2729.2009.00316.x

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16*(5), 367–371. https://doi.org/10.1111/j.0956-7976.2005.01542.x

Jipson, J. L., & Gelman, S. A. (2007). Robots and rodents: Children's inferences about living and nonliving kinds. *Child Development*, *78*(6), 1675–1688. https://doi.org/10.1111/j.1467-8624.2007.01095.x

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Human-Computer Interaction*, *19*(1), 61–84. https://doi.org/10.1207/s15327051hci1901&2_4

Kanda, T., Shimada, M., & Koizumi, S. (2012). Children learning with a social robot. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12* (p. 351). New York, New York, USA: ACM Press. https://doi.org/10.1145/2157689.2157809

Kazakoff, E. R., Sullivan, A., & Bers, M. U. (2013). The Effect of a Classroom-Based Intensive Robotics and Programming Workshop on Sequencing Ability in Early Childhood. *Early Childhood Education Journal*, *41*(4), 245–255. https://doi.org/10.1007/s10643-012-0554-5

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, *24*(2), 313–334. https://doi.org/10.1080/01690960802365567

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511807572

Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., … Belpaeme, T. (2017). Child Speech Recognition in Human-Robot Interaction. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17* (pp. 82–90). New York, New York, USA: ACM Press. https://doi.org/10.1145/2909824.3020229

Kita, S. (2009). Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes*, *24*(2), 145–167. https://doi.org/10.1080/01690960802586188

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

Kory Westlund, J. M., Jeong, S., Park, H. W., Ronfard, S., Adhikari, A., Harris, P. L., … Breazeal, C. L. (2017). Flat vs. Expressive Storytelling: Young Children's Learning and Retention of a Social

Robot's Narrative. *Frontiers in Human Neuroscience*, *11*, 295. https://doi.org/10.3389/fnhum.2017.00295

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196. https://doi.org/10.1017/S0272263116000085

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159. https://doi.org/10.2307/2529310

Lemaignan, S., Fink, J., & Dillenbourg, P. (2014). The dynamics of anthropomorphism in robotics. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14* (pp. 226–227). New York, New York, USA: ACM Press. https://doi.org/10.1145/2559636.2559814

Levine, J., & Resnick, L. B. (1993). Social Foundations of Cognition. *Annual Review of Psychology*, *44*(1), 585–612. https://doi.org/10.1146/annurev.psych.44.1.585

Li, D., Rau, P. L. P., & Li, Y. (2010). A Cross-cultural Study: Effect of Robot Appearance and Task. *International Journal of Social Robotics*, *2*(2), 175–186. https://doi.org/10.1007/s12369-010-0056-9

Liu, M., Horton, L., Olmanson, J., & Toprac, P. (2011). A study of learning and motivation in a new media enriched environment for middle school science. *Educational Technology Research and Development*, *59*(2), 249–265. https://doi.org/10.1007/s11423-011-9192-7

Macedonia, M. (2014). Bringing back the body into the mind: Gestures enhance word learning in foreign language. *Frontiers in Psychology*, *5*(DEC), 1467. https://doi.org/10.3389/fpsyg.2014.01467

Macedonia, M., & Knösche, T. R. (2011). Body in mind: How gestures empower foreign language learning. *Mind, Brain, and Education*, *5*(4), 196–211. https://doi.org/10.1111/j.1751-228X.2011.01129.x

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping*, *32*(6), 982–998. https://doi.org/10.1002/hbm.21084

Macedonia, M., & Von Kriegstein, K. (2012). Gestures Enhance Foreign Language Learning. *Biolinguistics*, *6*(3–4), 393–416. Retrieved from http://www.biolinguistics.eu

Marton, F., & Booth, S. (1997). *Learning and Awareness*. New York: Routledge. https://doi.org/10.4324/9780203053690

Masson-Carro, I., Goudbeek, M., & Krahmer, E. J. (2016). Can you handle this? The impact of object affordances on how co-speech gestures are produced. *Language, Cognition and Neuroscience*, *31*(3), 430–440. https://doi.org/10.1080/23273798.2015.1108448

Matuszek, C. (2018). Grounded language learning: Where robotics and NLP meet. In *IJCAI International Joint Conference on Artificial Intelligence* (Vol. 2018-July, pp. 5687–5691). California: International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2018/810

Mavilidi, M. F., Okely, A. D., Chandler, P., Cliff, D. P., & Paas, F. (2015). Effects of integrated physical exercises and gestures on preschool children's foreign language vocabulary learning. *Educational Psychology Review*, *27*(3), 413–426. https://doi.org/10.1007/s10648-015-9337-z

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Mazzoni, E., & Benvenuti, M. (2015). A robot-partner for preschool children learning English using Socio-Cognitive conflict. *Educational Technology and Society*, *18*(4), 474–485. https://doi.org/10.2307/jeductechsoci.18.4.474

McGregor, K. K. (2008). Gesture supports children's word learning. *International Journal of Speech-Language Pathology*, *10*(3), 112–117. https://doi.org/10.1080/17549500801905622

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23092060

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought* (1st ed.). Chicago: University of Chicago Press.

Mellor, D., & Moore, K. A. (2014). The use of likert scales with children. *Journal of Pediatric Psychology*, *39*(3), 369–379. https://doi.org/10.1093/jpepsy/jst079

Michalowski, M. P., Sabanovic, S., & Simmons, R. (2006). A spatial model of engagement for a social robot. In *International Workshop on Advanced Motion Control, AMC* (Vol. 2006, pp. 762–767). IEEE. https://doi.org/10.1109/AMC.2006.1631755

Mittelberg, I., & Evola, V. (2015). 131. Iconic and representational gestures. In C. Müller, A. Cienki, E. Fricke, S. Ladewig, D. McNeill, & J. Bressem (Eds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 38/2* (pp. 1732–1746). Berlin, München, Boston: De Gruyter. https://doi.org/10.1515/9783110302028.1732

Morett, L. M., Gibbs, R. W., & Macwhinney, B. (2012). The Role of Gesture in Second Language Learning: Communication, Acquisition, & Retention. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*(34). Retrieved from https://mindmodeling.org/cogsci2012/papers/0143/paper0143.pdf

Movellan, J. R., Eckhardt, M., Virnes, M., & Rodriguez, A. (2009). Sociable robot improves toddler vocabulary skills. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction - HRI '09* (p. 307). New York, New York, USA: ACM Press.

https://doi.org/10.1145/1514095.1514189

Mubin, O., Stevens, C. J., Shahid, S., Mahmud, A. Al, & Dong, J.-J. (2013). A Review of the Applicability of Robots in Education. *Technology for Education and Learning*, *1*(1). https://doi.org/10.2316/journal.209.2013.1.209-0015

Müller, C. (1998). Iconicity and gesture. In *Oralité et gestualité: Communication multimodale et interaction* (Serge Sant, pp. 321–328). Montréal/Paris: L'Harmattan.

Mumford, K. H., & Kita, S. (2014). Children Use Gesture to Interpret Novel Verb Meanings. *Child Development*, *85*(3), 1181–1189. https://doi.org/10.1111/cdev.12188

Novack, M. A., Goldin-Meadow, S., & Woodward, A. L. (2015). Learning from gesture: How early does it happen? *Cognition*, *142*, 138–147. https://doi.org/10.1016/j.cognition.2015.05.018

O'Brien, H. L., & Toms, E. G. (2008). What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, *59*(6), 938–955. https://doi.org/10.1002/asi.20801

Özçalişkan, Ş., & Goldin-Meadow, S. (2005). Gesture is at the cutting edge of early language development. *Cognition*, *96*(3), 101–113. https://doi.org/10.1016/j.cognition.2005.01.001

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database. In *ACM/IEEE International Conference on Human-Robot Interaction* (pp. 105–113). https://doi.org/10.1145/3171221.3171268

Piaget, J. (1952). *The origins of intelligence in children. American Journal of Psychotherapy* (Vol. 8). New York: W W Norton & Co. https://doi.org/10.1037/11494-000

Ping, R., & Goldin-Meadow, S. (2010). Gesturing Saves Cognitive Resources When Talking About Nonpresent Objects. *Cognitive Science*, *34*(4), 602–619. https://doi.org/10.1111/j.1551-6709.2010.01102.x

Post, L. S., Van Gog, T., Paas, F., & Zwaan, R. A. (2013). Effects of simultaneously observing and making gestures while studying grammar animations on cognitive load and learning. *Computers in Human Behavior*, *29*(4), 1450–1455. https://doi.org/10.1016/j.chb.2013.01.005

Ray, C., Mondada, F., & Siegwart, R. (2008). What do people expect from robots? In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS* (pp. 3816–3821). IEEE. https://doi.org/10.1109/IROS.2008.4650714

Riek, L. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119–136. https://doi.org/10.5898/jhri.1.1.riek

Rizzolatti, G., & Craighero, L. (2004). The Mirror-Neuron System. *Annual Review of Neuroscience*, *27*(1),

169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*(9), 661–670. https://doi.org/10.1038/35090060

Ros, R., Baroni, I., & Demiris, Y. (2014). Adaptive human-robot interaction in sensorimotor task instruction: From human to robot dance tutors. *Robotics and Autonomous Systems*, *62*(6), 707–720. https://doi.org/10.1016/j.robot.2014.03.005

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain SES disparities in child vocabulary size at school entry. *Science*, *323*(5916), 951–953. https://doi.org/10.1126/science.1167025

Rudovic, O., Lee, J., Mascarell-Maricic, L., Schuller, B. W., & Picard, R. W. (2017). Measuring Engagement in Robot-Assisted Autism Therapy: A Cross-Cultural Study. *Frontiers in Robotics and AI*, *4*, 36. https://doi.org/10.3389/frobt.2017.00036

Ruiz-del-Solar, J., & Avilés, R. (2004). Robotics courses for children as a motivation tool: The Chilean experience. *IEEE Transactions on Education*, *47*(4), 474–480. https://doi.org/10.1109/TE.2004.825063

Sabanovic, S., Michalowski, M. P., & Simmons, R. (2006). Robots in the wild: Observing human-robot social interaction outside the lab. In *International Workshop on Advanced Motion Control, AMC* (Vol. 2006, pp. 576–581). IEEE. https://doi.org/10.1109/AMC.2006.1631758

Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2013). To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability. *International Journal of Social Robotics*, *5*(3), 313–323. https://doi.org/10.1007/s12369-013-0196-9

Sawyer, R. K. (2014). *The Cambridge handbook of the learning sciences, second edition*. (R. K. Sawyer, Ed.), *The Cambridge Handbook of the Learning Sciences, Second Edition*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139519526

Schneider, S., Häßler, A., Habermeyer, T., Beege, M., & Rey, G. D. (2019). The more human, the higher the performance? Examining the effects of anthropomorphism on learning with media. *Journal of Educational Psychology*, *111*(1), 57–72. https://doi.org/10.1037/edu0000273

Shatz, I. (2016). Native Language Influence During Second Language Acquisition: A Large-Scale Learner Corpus Analysis.

Shields, B. J., Palermo, T. M., Powers, J. D., Grewe, S. D., & Smith, G. A. (2003). Predictors of a child's ability to use a visual analogue scale. *Child: Care, Health and Development*, *29*(4), 281–290. https://doi.org/10.1046/j.1365-2214.2003.00343.x

Shimada, M., Kanda, T., & Koizumi, S. (2012). How can a social robot facilitate children's collaboration? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 7621 LNAI, pp. 98–107). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-34103-8_10

Sidner, C. L., Kidd, C. D., Lee, C., & Lesh, N. (2004). Where to look: A study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interface - IUI '04* (p. 78). New York, New York, USA: ACM Press. https://doi.org/10.1145/964442.964458

Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., & Rich, C. (2005). Explorations in engagement for humans and robots. *Artificial Intelligence*, *166*(1–2), 140–164. https://doi.org/10.1016/j.artint.2005.03.005

Slangen, L., Van Keulen, H., & Gravemeijer, K. (2011). What pupils can learn from working with robotic direct manipulation environments. *International Journal of Technology and Design Education*, *21*(4), 449–469. https://doi.org/10.1007/s10798-010-9130-8

SoftBank Robotics. (n.d.). Joints — Aldebaran 2.1.4.13 documentation. Retrieved July 9, 2019, from http://doc.aldebaran.com/2-1/family/robots/joints_robot.html

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–699. https://doi.org/10.1111/j.0023-8333.2005.00320.x

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7Get

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. *Cognitive Load Theory*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4419-8126-4

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, *104*(46), 17954–17958. https://doi.org/10.1073/pnas.0707769104

Tanaka, F., & Matsuzoe, S. (2012). Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning. *Journal of Human-Robot Interaction*, 78–95. https://doi.org/10.5898/jhri.1.1.tanaka

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *8*(2), 219–235. https://doi.org/10.1075/gest.8.2.06tel

Tung, F.-W. (2016). Child Perception of Humanoid Robot Appearance and Behavior. *International Journal of Human-Computer Interaction*, *32*(6), 493–502.

https://doi.org/10.1080/10447318.2016.1172808

Ültanır, E. (2012). An Epistemological Glance at the Constructivist Appraoch: Constructivist Learning in Dewey, Piaget, and Montessori. *International Journal of Instruction*, *5*(2), 196–212.

van den Berghe, R., de Haas, M., Oudgenoeg-Paz, O., Krahmer, E. J., Verhagen, J., Vogt, P., … Leseman, P. (2019). A toy or a friend? Children's anthropomorphistic beliefs about robots and the relation with second language word learning. *(In Preparation)*. Retrieved from http://www.l2tor.eu/effe/wp-content/uploads/2015/12/Proefschrift-Rianne-van-den-Berghe.pdf

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

van Dijk, E. T., Torta, E., & Cuijpers, R. H. (2013). Effects of Eye Contact and Iconic Gestures on Message Retention in Human-Robot Interaction. *International Journal of Social Robotics*, *5*(4), 491–501. https://doi.org/10.1007/s12369-013-0214-y

van Elk, W. J. (2018). *De leermiddelenketen in het onderwijs*. Zoetermeer: Kennisnet. Retrieved from https://www.kennisnet.nl/fileadmin/kennisnet/publicatie/Kennisnet-leermiddelenketen-in-het-onderwijs.pdf

van Nispen, K., van de Sandt-Koenderman, M., Mol, L., & Krahmer, E. J. (2014). Pantomime Strategies: On Regularities in How People Translate Mental Representations into the Gesture Modality. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society (CogSci 2014)* (pp. 3020–3025). Austin, Texas: Cognitive Science Society. Retrieved from https://pure.uvt.nl/ws/portalfiles/portal/4314699/paper521.pdf

Varney, M. W., Janoudi, A., Aslam, D. M., & Graham, D. (2012). Building young engineers: TASEM for third graders in woodcreek magnet elementary school. *IEEE Transactions on Education*, *55*(1), 78–82. https://doi.org/10.1109/TE.2011.2131143

Vogt, P., de Haas, M., de Jong, C., Baxter, P., & Krahmer, E. J. (2017). Child-Robot Interactions for Second Language Tutoring to Preschool Children. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00073

Vogt, P., Van Den Berghe, R., De Haas, M., Hoffman, L., Kanero, J., Mamus, E., … Pandey, A. K. (2019). Second Language Tutoring Using Social Robots: A Large-Scale Study. In *ACM/IEEE International Conference on Human-Robot Interaction* (Vol. 2019-March, pp. 497–505). IEEE. https://doi.org/10.1109/HRI.2019.8673077

Washington State Department of Social & Health Services. (2014). *Range of Joint Motion Evaluation Chart*. Retrieved from https://www.dshs.wa.gov/sites/default/files/FSA/forms/pdf/13-585a.pdf

Westlund, J. M. K., Martinez, M., Archie, M., Das, M., & Breazeal, C. L. (2016). Effects of framing a robot as a social agent or as a machine on children's social behavior. In *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016* (pp. 688–693). IEEE. https://doi.org/10.1109/ROMAN.2016.7745193

Whittier, L. E., & Robinson, M. (2007). Teaching Evolution to Non-English Proficient Students by Using Lego Robotics. *American Secondary Education*, *35*(3), 19–28. https://doi.org/10.2307/41406087

Williams, D. C., Ma, Y., Prejean, L., Ford, M. J., & Lai, G. (2007). Acquisition of physics content knowledge and scientific inquiry skills in a robotics summer camp. *Journal of Research on Technology in Education*, *40*(2), 201–216. https://doi.org/10.1080/15391523.2007.10782505

Wyer, R. S. (2007). Principles of mental representation. In A. W. Kruglanski, & E. T. Higgins (Eds.). In *Social psychology: Handbook of basic principles* (pp. 285–307). New York: Guilford Press.

Appendices

Appendix A – List and Description of All Used Gestures – Including Videos

Table A1. List of all used gestures in this study

Gesture Name	Description	Video Link
*Bridge 1**	Two arms form bridge, hand 'walks' across	https://youtu.be/zGhDmzUpc1k
Bridge 2	Arms wide	https://youtu.be/pR941oLLCPw
Bridge 3	Hunches over, forms arch with arms and body	https://youtu.be/c3WS9RCpoHY
Bridge 4	Uses two arms to make wide line; hand walks across	https://youtu.be/5R6HbsCILpU
Bridge 5	Uses one arm to form line; two arch motions with hand	https://youtu.be/USnsYzq7A4U
Horse 1	Slow grazing motion	https://youtu.be/Dkf5LJA9AyQ
Horse 2	Slow Horse-riding motion	https://youtu.be/AMTHYpHe5oc
*Horse 3**	Fast horse-riding motion with lasso	https://youtu.be/3T9n2q2OLZk
Horse 4	Prancing horse motion with arms	https://youtu.be/EOXAhDS7Ct0
Horse 5	Fast horse-riding motion	https://youtu.be/VuJsqsrsM3g
Pencil 1	Hunched over writing motion	https://youtu.be/bpLuYYPFDOw
Pencil 2	Writing in hand motion	https://youtu.be/yl2Xw6GE4As
Pencil 3	Air writing motion	https://youtu.be/2XfJa2-1GXI
*Pencil 4**	Two arms signifying long object; air writing motion	https://youtu.be/82PKmaBydDY
Pencil 5	Two arms in air, signifying pencil head	https://youtu.be/A3ulHlEHdxk
*Spoon 1**	Using arm to dip [in container]; spoon-eating motion	https://youtu.be/IMyoJznkb-s
Spoon 2	Transferring between two containers motion	https://youtu.be/umQGkzKBnto
Spoon 3	Spoon shape motion with two arms; eating motion	https://youtu.be/zseXPJuhAic
Spoon 4	Stirring motion	https://youtu.be/iQFqjZDsry4
Spoon 5	Eating from container in hand	https://youtu.be/y6jm8JTWOGA
Stairs 1	Making stair shape in air	https://youtu.be/4BZFgGkb6PI
Stairs 2	'Ollekebolleke' motion (typical Dutch; placing fist on fist signifying height)	https://youtu.be/VYK1LT73dss
Stairs 3	Legs stepping; hand going into air	https://youtu.be/8hLS2XciVeY
Stairs 4	Hand on top of hand	https://youtu.be/irwPIbeCeCo
*Stairs 5**	Ladder climbing motion	https://youtu.be/g5FxYIPzKdU
Turtle 1	Slow swimming motion	https://youtu.be/inHxlfsPaNU
Turtle 2	Shield above head	https://youtu.be/e2WCrJYlpG0
*Turtle 3**	Peeking out of shield motion	https://youtu.be/j9M_iTxeLvE
Turtle 4	Slow walking motion	https://youtu.be/_geRQu70Bl0
Turtle 5	Side-ways pointing at shield on back	https://youtu.be/GrXvRHSnceg

Notes. Asterisks denote best scoring gesture in perception study, this gesture was subsequently used in the single-gesture condition.

A full playlist of all videos can be watched here:

https://www.youtube.com/watch?v=GrXvRHSnceg&list=PLv8YzCSmJE7BDEXVD1k7D2j9q2MGeim67

Appendix B – Information Letter and Consent Form for Schools and Parents (Dutch)

Sociale Robots
als tweede taal-leraar

Beste ouder(s)/verzorger(s),

Vanuit *Tilburg University* wordt onderzoek gedaan naar het gebruik van sociale robots om jonge kinderen een tweede taal te leren. De school van uw kind heeft aangegeven mee te willen werken aan dit onderzoek. Wij willen u via deze brief informeren over het onderzoek en vragen om uw medewerking.

School of Humanties and Digital Sciences

TILBURG UNIVERSITY

Understanding Society

Tilburg University

Tilburg, mei 2019

Doel van het onderzoek

Het onderzoek, dat in mei van start zal gaan, richt zich op het ontwikkelen van sociale robots die aan kleuters een tweede taal kunnen leren. Het beheersen van een tweede taal wordt steeds belangrijker met de toenemende internationalisering. Het is bekend dat hoe jonger kinderen beginnen met het aanleren van een tweede taal, hoe beter zij die taal later zullen beheersen. Nu is het echter vaak het geval dat kinderen pas op oudere leeftijd een tweede taal leren. Daarnaast is gebleken dat kinderen met één op één lessen meer taal leren. Op school heeft de leerkracht niet altijd tijd om kinderen één op één te helpen een tweede taal te leren. Maar voor een robot is dat geen probleem.

De resultaten van het onderzoek leveren belangrijke informatie op over de sociale en talige vaardigheden die de robot moet beheersen om tweede tass te kunnen geven aan jonge kinderen. Deze kennis over de benodigde vaardigheden zal bijdragen aan de ontwikkeling van een robot die niet alleen sociaal is, maar ook kan helpen bij het lesgeven.

Waar bestaat deelname uit?

Als uw kind meedoet aan het onderzoek, zal hij /zij na een introductie van de robot in kleine groepen, eenmalig met de robot een spelletje spelen op een tablet computer. Tijdens dit spel vraagt de robot in een speelse interactie om objecten van Engelse woordjes aan te wijzen. Het doel is om uw kind via deze weg spelenderwijs de Engelse woorden te leren. Dit spel duurt ongeveer 30 minuten. Vlak voor en na de les sessie zullen we een korte test doen om te kijken welke woorden uw kind heeft geleerd, en deze test wordt na een week nogmaals herhaald. De testen duren maximaal 5 minuten. Verder zullen de kinderen een aantal korte vragen krijgen over wat zij van de robot vinden (bijvoorbeeld: "Denk je dat de robot het voelt als je hem kietelt?"). Het onderzoek zal onder schooltijd plaatsvinden op een tijdstip dat in overleg met de leerkracht wordt vastgesteld. Tijdens de taak zal uw kind gefilmd worden zodat wij daarna, op basis van de beelden, kunnen onderzoeken hoe uw kind met de robot omgaat. Te allen tijde zal er een experimenteerleider aanwezig zijn. Als u toestemming geeft om uw kind mee te laten doen, dan doen we dat alleen als uw kind dat ook wilt. De meeste kinderen vinden het leuk om met de robot te spelen. Maar als uw kind voor of tijdens het onderzoek op de een of andere manier aangeeft niet meer mee te willen doen, zullen we het onderzoek met uw kind direct stopzetten. Hier zijn geen gevolgen aan verbonden. De veiligheid en welzijn van uw kind hebben bij ons altijd eerste prioriteit.

Wist u dat...?

Er wordt op het moment veel onderzoek gedaan na de effectiviteit van robots in de onderwijswereld. Dit is dan ook niet het eerste onderzoek dat Tilburg University hiernaar doet. Eind 2018 is een grootschalig onderzoek ten einde gekomen, genaamd L2TOR. Het huidige onderzoek is hier een vervolg op.

Wilt u meer weten? Kijk dan gerust eens op de website van L2TOR. Hier vindt u tevens een video welk u een beeld geeft hoe de kinderen met de robot om gaan.

www.l2tor.eu

Wat gebeurt er met de videobeelden en andere gegevens?

De beelden en andere gegevens (zoals de leeftijd en het geslacht van uw kind, en het aantal woorden dat ze hebben geleerd) zullen alleen gebruikt worden voor onderzoeksdoeleinden en niet voor enig ander doel zonder uw toestemming. Na afloop van het onderzoek zullen de beelden en gegevens nog 10 jaar op een beveiligde computer op *Tilburg University* worden bewaard, waarna ze vernietigd worden. Uiteraard worden de gegevens van uw kind anoniem verwerkt, en zal de naam van uw kind los van de andere gegevens worden bewaard. De gegevens worden bewaard volgens de richtlijnen van de Algemene Verordening Gegevensbescherming, en u heeft recht om te allen tijde de gegevens in te zien en door ons te laten verwijderen. Het onderzoek wordt uitgevoerd met toestemming van de ethische onderzoekscommissie van de *Tilburg School of Humanities and Digital Sciences*. Na afloop van het onderzoek zullen de algemene resultaten middels een brief aan u worden teruggekoppeld.

Wij hopen van harte dat uw kind en u mee willen doen aan ons onderzoek. Deelname is vrijwillig en er zijn geen nadelige consequenties verbonden aan deelname, maar ook geen directe voordelen voor uw kind of uzelf. Indien u toestemming geeft voor deelname, vragen wij u het bijgevoegde toestemmingsformulier in te vullen, te ondertekenen en aan uw kind mee te geven naar school. Bij vragen kunt u contact opnemen met ondergetekenden of de website www.l2tor.eu bezoeken. Voor eventuele vragen en opmerkingen tijdens of na afloop van het onderzoek, kunt u contact opnemen met een van de onderzoekers (Arold Brandse, of Jan de Wit

Bij voorbaat dank.

Met vriendelijke groet,
Arold Brandse
Jan de Wit (projectleider)

Tilburg University

ID-nummer

Toestemmingsformulier onderzoek taalles

- Ik bevestig dat ik de informatiebrief voor ouders heb gelezen. Ik heb de gelegenheid gehad om aanvullende vragen te stellen. Deze vragen zijn in voldoende mate beantwoord.
- Ik heb voldoende tijd gehad om over deelname na te denken.
- Ik weet dat deelname aan het onderzoek door mijn kind geheel vrijwillig is en dat ik mijn toestemming op ieder moment kan intrekken zonder daarvoor een reden op te geven. Hier zijn geen gevolgen aan verbonden.
- Ik geef toestemming om de gegevens van mijn kind te verwerken voor de doeleinden zoals beschreven in de informatiebrief.
- Ik heb geen bezwaar tegen het opnemen van mijn kind op video voor onderzoeksdoeleinden. Ik weet dat de opnames door geen anderen dan de bij het wetenschappelijk onderzoek of behandeling betrokkenen bekeken zullen worden
  - ☐ (Optioneel, vink dit alstublieft aan als u toestemming geeft). Ik geef toestemming dat anoniem materiaal (bijvoorbeeld een foto) van mijn kind gebruikt mag worden in een wetenschappelijke publicatie of presentatie (het gezicht van uw kind zal onherkenbaar gemaakt zijn).
- Ik geef toestemming om de gegevens van mijn kind die zijn verkregen met dit onderzoek gedurende 10 jaar na afloop van het onderzoek te bewaren.
- Ik stem in met deelname van mijn kind en ons als ouders/verzorgers aan bovengenoemd onderzoek.
  - ☐ (Optioneel, vink dit aan wanneer u dit wilt) Ik wil graag na afloop van het onderzoek een brief met het onderzoeksresultaat ontvangen. Mijn emailadres is _____

Gegevens van uw kind

Voor- en achternaam kind _____

Geboorte datum _____

Moedertaal kind _____

Spreekt uw kind meer talen? *Ja / Nee (zo ja, welke?)* _____

Spreekt u thuis meer talen? *Ja / Nee (zo ja, welke?)* _____

Handtekening ouder(s) / verzorger(s)

_____

Naam ouder(s) / verzorger(s) _____

Datum _____

Tilburg University

ID-nummer

In te vullen door lid van het onderzoeksteam

Ondergetekende verklaart dat de hierboven genoemde ouder/verzorger schriftelijk over het bovengemelde onderzoek is geïnformeerd. Hij/zij verklaart tevens dat deelname door bovengenoemd kind, van geen enkele invloed zal zijn op het onderwijs dat hem/haar toekomt. Alle gegevens zullen strikt vertrouwelijk worden behandeld. De verwijzingen die de identiteit van de ouder/verzorger of kind kunnen onthullen, zullen zorgvuldig worden verwijderd.

Handtekening onderzoeksteam

Naam _____

Functie _____

Datum _____

Appendix C – Coding Book Engagement Videos

Table C1. *Codebook for Child-Task Engagement. Based on de Haas et al. (2019)*

-2	• No concentration: staring and dreaming • Completely absent and passive attitude • No specific activity; purposeless acts • Focusing on experiment leader instead of task • No signs of exploration or interest • No mental activity
-1	• Limited concentration: looking away, fiddling, short dreaming • Easily distracted • Limited execution of task • Starts fiddling with tablet
0	• Child is doing task, but routinely, fleetingly, without putting too much thought into the answers • Limited motivation, is not challenged, does not show enthusiasm • Is not gaining profound experiences • Is not fully absorbed in task • Child uses moderate amount of mental capacity • Most tasks are done without intervention
+1	• Is mostly absorbed by task; (softly) repeats words sometimes • Moderate concentration, sometimes weakening • Child feels challenged, driven by task; shows that they are actively trying to look for the correct answer (hovering over multiple answers) • Uses mental capacity • Appeals to the child's imagination and cognitive capacity
+2	• Child is fully concentrated on task; absorbed in task; repeats words often • Shows excessive motivation to complete task; tries to always make sure they are giving the correct answer • Cannot be distracted • Has attention for details • Uses full mental capacity and abilities • Enjoys being driven

Note. When looking at the robot when it speaks or performs a gesture, this does not mean lower child-task engagement. Only if the child focuses on something that is unrelated to the robot (when it should be) or task, is it scoring lower on child-task engagement.

Table C2. *Codebook for Child-Robot Engagement. Based on de Haas et al. (2019)*

-2	• Ignores the robot completely • Has a closed body posture towards the robot • Completely absent and passive attitude • No specific activity; purposeless acts • Focusing on experiment leader instead of task • No signs of exploration or interest • No mental activity
-1	• Limited concentration: looking away, fiddling, short dreaming • Limited amounts of looking at robot (especially when robot is merely talking); not even when performing gestures • Easily distracted • Limited execution of task
0	• Limited motivation, is not challenged, does not show enthusiasm • Is not gaining profound experiences • Child uses moderate amount of mental capacity • In gesture condition: generally only watches when robot performs a gesture; sometimes very short intervals at robot in between
+1	• Sometimes speaks to robot • Actively looks at robot when it is speaking • Is mostly absorbed by robot • Moderate concentration, sometimes weakening • Explores what the robot is doing; shows a (visible) reaction to a gesture for example • Appeals to the child's imagination and cognitive capacity
+2	• Actively speaks to robot • Child is fully concentrated on robot; absorbed in robot • Shows excessive interest in robot • Has much attention for robot; looks at details • Mimics gestures • Uses full mental capacity and abilities • Enjoys being driven

Note. If child looks at robot while it repeats words, this counts for child-robot engagement as well. Merely repeating words while not looking at the robot does not count as child-robot engagement.

Table C3. Codebook for Valence. Based on Rudovic et al. (2017)

-2	• Severe unpleasant feelings (unhappy, angry, upset, frightened) • Severe disappointment • Miserable, annoyed • Emotions are very visible and possibly audible in child
-1	• Lower intensity negative emotion; • No audible display of emotion; mostly visual
0	• Neutral emotion; no display of emotion
+1	• Lower intensity positive emotion; • Sometimes smiles (when getting a correct answer)
+2	• Intense happiness (e.g. clapping of hands) • Intense Joy (e.g. with audible laughter) • Intense delight

Table C4. Codebook for Arousal. Based on Rudovic et al. (2017)

-2	• Extremely bored, walking away from interaction • Extremely sleepy; excessive yawning • Very passive stance towards robot and task • Sighs often
-1	• Slightly bored • Starts thinking of other ways to make correct answer (using elbows, weird hand constructions, etc.) • Rests on arms • Sometimes yawns • Rubbing into eyes • Hanging back into chair • Starts sighing
0	• Neutral arousal • No excitement nor boredom visible
+1	• Slightly aroused • Visibly thinks of correct answer • Shows active signs when receiving feedback on right or wrong answers
+2	• Highly aroused • Very active and spontaneous stance towards task and robot • Active bodily movements (related to task and / or robot)

Appendix D – Protocol Written for Study

Protocol
Variation study

May, 2019

School of Humanties and Digital Sciences

TILBURG UNIVERSITY

Understanding Society

TILBURG ◆ ◆ UNIVERSITY

Checklist equipment

- NAO Robot
- NAO Charger
- Surface Pro 4
- Surface Pro 4 charger
- 1 Laptop (For tests and as control panel)
- Router
- 2 x Video camera
- 2 x SD card (32+ GB)
- Extension cord
- Ethernet cable x3 (robot to router, tablet to router, laptop to router)
- Mouse (we have two special kids mice in D329)
- Plastic container (to put the tablet on)
- Papercraft pages (as present after post-test)
- Consent forms
- Pen and paper
- Perception Study Questionnaire papers
- Wipes (for cleaning tablet / laptop)
- "Do not enter" sign

Setup of the experiment

Note that there are several different conditions for the current experiment:

1. No Gesture condition: The robot will not perform any gestures, will use only speech for target words.
2. Single Gesture condition: The robot will perform the same iconic gesture each time a target word is mentioned.
3. Varied Gestures condition: The robot will perform a different varation of a gesture each time a target word is mentioned.

Stages of experiment

Introduction

The introduction is the first time the children will see and interact with the robot. Its main funtion is to alleviate any fears the children may have towards something they have no experience with yet, a robot. During this introduction, the children will start will developing a relationship with the robot, mainly as being a like-minded but slightly different personification. Robin the robot will be introduced as a 7 year old entity who is planning on visiting the United Kingdom in the near future, the reason for his willingness to learn the English language.

Ideally, the introduction is done with two researchers present. One can interact with the robot and the children, while the other control the robot using the controlpanel (this second person should try to remain inconspicuous). Before the robot does its programmed routine, he should be introduced to the children. A sample text of this can be found in Appendix A. While some deviation is allowed, depending on how much time is available and other variables, the main points should always be told: Robin is 7 years old. He is a robot who, while looking like us, is slightly different than us.

After the introductory text, the control panel available on the tablet can be used to allow the robot to follow its routine. Using the enter button, the robot can continue to the next part of its introduction. Please take note that attempting to skip certain parts is not recommended; the robot will start combining different animations and may potentially launch itself into the air. After the robot has done its first dance routine ("Hoofd, schouders, knie en teen"), the robots arm is set into a 'free' mode. During this time, children who wish to do so, can shake hands with the robot. Make sure to keep the robot on your own lap and assist with holding his arm, as children may be fairly aggressive in their interaction with the robot. Again, the main focus of this introduction is to alleviate any fears in the children; encourage them to interact with the robot but do not push them if they do not wish to do so. After the introduction is done, the reseachers can commence with pre-testing.

Pre-test

After the initial introduction, children who have handed in the experiment consent forms (filled in and approved by their parents) can be pre-tested on their Dutch and English knowledge. In consultation with the teachers, the researchers can get a single participant and administer the pre-tests.

As the children have just met with Robin, it is mostly ideal to start with the perception study questionnaire (Appendix B). Fill in the date and the ID-number (corresponding with the later Log files of the pre- and post-tests). Tell them that you are curious about what they think of Robin. The initial answers of the test (yes / no / I don't know) will be tallied and compared, so these are the most important. But always try to ask shortly what their reasoning is and note this on the paper.

Following the perception study questionnaire, the Dutch and English word knowledge pre-tests will be held. These test will be available on the accompanied laptop. Ask the child if they are open to playing a short game (they most likely will be), and that after the game they are done and can go back to their class. For consistency, it is vital that each pre-test is taken in a similar manner:

Before the child enters the room, have the pre-test ready to run on the laptop. Start the pre-test file ("1 pre-tests NL and EN.bat"). Once the browser has started up, press F11 (sometimes FN has to be pressed too) to place the browser in full-screen mode. The initial screen should be a green smiley.

Ask the child to sit in front of the laptop on a chair or stool. Sit beside them and tell them they are going to play a short language game ("taalspelletje"). For the first test, explain that the computer will mention a Dutch word of an object, after which they will be presented with six images. They have to choose the correct image that corresponds with the word. Then tell them that if they understand, they can tap the green smiley to start the game. During the game, if the child is taking very long to chose an answer (and is not visibly doubting between multiple answers), encourage them by telling them to pick one of the options. If a child mentions that he did not hear the word correctly, you can press the "R" on the keyboard to have the computer mention the word once more. This Dutch pre-test tests for six words, once.

When the screen displays a thumbs-up icon, with confetti behind it, the Dutch portion of the test is done. Pressing the "Space" key at this point allows you to continue with the English pre-test. Explain to the child that they are now going to do a similar game but with English words. It is vital here to let the child know that they are not expected to know the answer. Tell them to just chose which object they find fitting best with the mentioned word. Again, they can press the green smiley if they understand and can begin with the pre-test. This English pre-test tests for six words, three times with different images. After the English pre-test is done, the child is free to go back to the class. Accompany them and retrieve the next child.

Backup: After the day has concluded, backup all log files to the Google Drive folder that has been set up for this experiment. The logs can be found in: www > logs > [file]. Do not delete the files from the laptops, as the results of the post tests are appended to this file. Make sure that the log files correspond with the ID-numbers on the consent form and the perception questionnaire form.

Training

Several days after the pre-tests, you will return for the training.

Setup

First, set up the robot, camera's, laptop and tablet according to figure 1. Connect all power cables. The connect all ethernet cables; one from the robot to the router, one from the control laptop to the router and one from the child's tablet to the router. Make sure that all cables are not in the way of any paths a child may take during entering of the room. Once done, turn on the robot by shortly pressing the button on it's chest. Note that it takes a relatively long time for the robot to start up. Furthermore, during set up of the camera's make sure that there is ample room for the child to be visible in the viewfinder. The child's face should be fully visible, with no parts of the child being blocked by the robot.

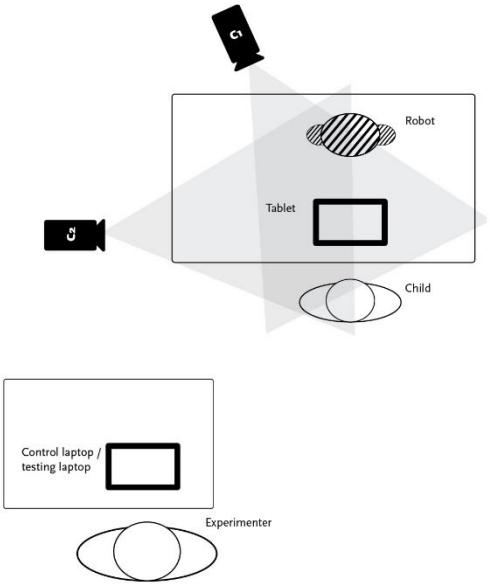


Figure 1. Training setup

Ideally, the tablet is placed on a table with a chair or stool in front of it, made specifically for the child's height. It is important for the child to sit comfortably for the duration of the training. The robot is placed directly behind the tablet. Place the tablet at a shallow angle so they can easily see the robot during the interaction (figure 2). On the tablet's desktop, you should find a folder called "Variatiestudie". Choose the correct file corresponding with the condition that you chose for the next child. Once the browser starts up, press F11 to make the browser fullscreen. Make sure to remove the keyboard neccessary for initiating the training program, prior to the child entering the room. Start up the concept-mapping file ("2 concept binding.bat") on the laptop and make the browser full-screen here as well.

Before retrieving a child for the training, make sure to turn on the camera's and have them start recording. Furthermore, test the names of the children in the control panel to make sure the robot announces the name correctly. If not, experiment with different (phonetic) spellings of the name. Finally, enter this name in the control panel, so it is ready for starting.

Figure 2. Tablet setup

Training

Once the set up is done, retrieve a child (in consultation with the teacher). Upon entering of the room, ask them if they are excited to start playing a game with Robin (most likely, they will be). In all cases, try to break the ice with the children. Ask them whether they can still remember how old Robin was (7 years). Then direct them towards the Robot and mention to them that Robin is a robot ("Robin is a robot, right?"). Then mention, that even though he is a robot, he has very similar features to use. Point towards the arm and ask them "Do you know what this is?". Do the same for the fingers and the robot's legs. Then conclude with saying that even though he is a robot, he looks quite a bit like us. At this point, some children will ask an additional question (e.g. "How does he breathe?"). Make up an answer mentioning a feature of Robin (e.g. "He is a robot, he uses electricity and doesn't need to breathe!"). Finally, tell them before starting the experiment, you want to play a short game on the laptop.

In a similar manner to the pre-test, sit beside the child facing the laptop. Explain to them that the computer will mention and English word and that the correct image will be shown on the screen. Instruct them to listen carefully, as they need this to play the game with Robin in a bit. Tell them to press the green smiley if they understand and start the concept mapping. Once done, tell them they can now join Robin to start playing a game with him. Mention that Robin will explain what game he wants to play himself. Furhtermore, explain to the child that you will busy doing something else while they are playing the game.

Once the child is seated, the start button can be pressed in the control panel (do not forget to add the child's name in the top field). For the next 20-30 minutes (time varies with conditions and the child's ability), the training will commence itself. If at any point something happens, refer to the "What to do if" section. If they happen to click the red smiley shown on the tablet after initial instructions by Robin because they have not understood the game, intervene. Try to explain how the game (I spy with my little eye..) works. When they eventually understand, have them click the green smiley shown on the tablet. Be seated behind and away from the child while they play with Robin.

During the training, note down any out-of-the-ordinary occurences of the interaction (Qualitative observations; e.g. the child turns around to ask if the game is nearly done; child starts to slouch; child mimics gestures, etc.). A round indicator should be present in the control panel to help with note taking.

Immediate Post-test

During the final phases of the training, set up the laptop for the post-test. Again, open up the corresponding file ("3 post-test EN.bat") and make the browser full-screen using F11.

Once the child is done with the training, join them and ask them if they enjoyed playing the game with Robin. During this short intermission, once more ask them the questions present in the perception study (similar to the pre-test; note: there is a slight difference between questions 12a and 12b).

Finally, after the questionnaire, instruct them that you once more want to play a short game with them on the laptop, and that they can return to class after playing this game. The post-test is similar in set up to playing the English pre-test. Use similar instructions to this.

Once done, turn of the camera's and accompany the child back to the class. Return to the experiment room and set up all devices for the next training. Note, if the child can do the post-test without any help from the researcher, the next experiment (tablet + condition choosing) can also be set up while the previous child is still doing the post-test.

Backup: After the day has concluded, backup all log files to the Google Drive folder that has been set up for this experiment. The logs can be found in: www > logs > [file]. Note that these results are added to the same file as the pre-test. Do not delete the files. Make sure that the log files correspond with the ID-numbers on the consent form and the perception questionnaire form. Make sure to also back-up all video files from the camera's and set them up for the next day.

Delayed Post-test
Approximately one week after the training sessions, you will once more return to the school to administer a delayed post-test. Set up the laptop similar to the pre-tests and open up the final file now ("4 delayed post-test EN.bat").

Finally, in consultation with the teacher, provide the children (as well as the children that were unable to participate) with the papercraft papers as an appreciation of their participation (Appendix C).

Backup: After the day has concluded, backup all log files to the Google Drive folder that has been set up for this experiment. The logs can be found in: www > retention > logs > [file]. Make sure that the log files correspond with the ID-numbers on the consent form and the perception questionnaire form.

What to do if

- ***The system fails prior to the interaction?***
  Reboot. If the experiment does not start, there may be something wrong with the IP address of the robot. You will have to update this both on the tablet and on the laptop running the control panel. On the tablet, right click the .bat file to start the experiment and choose "Edit". Then, change the IP address that is listed behind –i into the correct IP address – you can find out what the correct address is by pressing on the robot's chest button. On the desktop of the control panel laptop there should be a robotip.js file, you should also change the IP address there. Now, restart the experiment (follow the above steps again) and it should work!
- ***The system fails during the interaction?***
  Is it still the introduction? If so, restart the entire introduction (from the control panel).
- ***The child does not listen to instructions?***
  You may be stern if necessary (especially when they are likely to damage the equipment), for example by saying something along the lines of "<NAME OF CHILD>, ik wil dat je blijft zitten en met Robin gaat spelen.". If the situation becomes unmanageable, consult the teacher.
- ***The child is shy?***
  Try to break the ice: ask them context-relevant questions, briefly chit-chat, or encourage them to perform an action. Do not be too forward, but also do not be too reserved yourself. Make the child feel welcome and show that you are excited for them to play with the robot (reminder: make sure to refer to the robot as Robin when communicating with the child).
- ***The child is constantly seeking the attention of the experimenters?***
  Make it clear you are doing something else while they are interacting with the robot. If they ask you to provide them with the correct answer, don't. Do not even provide hints. Instead, simply tell them to pick whichever answer they like best/think is the correct answer (e.g., "Kies er maar één!").
- ***The child has to go to the bathroom?***
  Pause the experiment. At this age, the children should be able to go to and use the bathroom on their own. If this is not the case, consult their teacher.
- ***The child starts to cry?***
  Pause the experiment. Try to figure out why the child is crying and console if possible. If the child is inconsolable/continues to cry, consult the teacher. Stop the experiment if necessary.
- ***Something else happens?***
  If you cannot guarantee the safety/well-being of the child, stop the experiment immediately. The child's safety/well-being is our number one priority. However, if something happens that does not put the child in harm's way, intervene if possible (and deemed necessary) or discuss the appropriate course of action with your colleague/the child's teacher.
- ***Always make a note of irregularities, interruptions, and disruptions of the experiment!***

Variation Study - Appendix A: Sample introduction text (Dutch)

[optioneel] Voor de leerkracht:
Dit zijn [naam onderzoekers], en die hebben iets heel speciaals bij voor jullie. Ze gaan eerst een verhaaltje vertellen en dan gaan ze laten zien wat voor speciaals ze bij hebben.

NB: probeer te vermijden het geslacht van de robot te zeggen, dus geen "hij" maar "ie", "tie".

Dit is Robin,

(Laat foto zien)
Robin is niet helemaal zoals andere kindjes, want Robin is een robot. Hij woont pas net hier in [plaatsnaam], en komt speciaal voor jullie langs vandaag om eens te kijken bij jullie. Hij is net iets ouder dan jullie, want hij is 7 jaar. Robin wil graag nieuwe vriendjes maken, maar dat vindt 'ie nog wel een beetje eng, dus zullen we heel erg ons best doen om lief te zijn tegen Robin?

Omdat Robin een robot is ziet hij er wel wat anders uit dan wij allemaal. Maar Robin heeft wel armen net als wij hè? En wat heeft Robin nog meer zoals wij? Benen hè, ja klopt. En voetjes. En wat heeft Robin niet?

Als hij praat klinkt dat een beetje gek en beweegt zijn mond ook niet en hij kan jullie ook een beetje moeilijk verstaan, dus als jullie iets willen vragen aan Robin zullen wij wel antwoord geven. Ook al ziet ie er wat anders uit, Robin is wel heel aardig.

Robin gaat binnenkort op vakantie naar Engeland (foto Engeland). Weet iemand welke taal de mensen daar spreken? Engels inderdaad. En daarom is Robin nu een beetje Engels aan het leren. Robin vindt het leuk om te dansen op muziek, vinden jullie dat ook leuk? Houden jullie ook van muziek?

Robin is altijd heel erg blij na het dansen. Ik denk dat Robin jullie wel heel graag eens een keertje wil zien, dus zullen we hem maar eens roepen om te vragen of 'ie komt? Dan gaan we samen roepen, oké?

Je hoeft het niet spannend te vinden hoor, want alle kindjes zijn heel lief toch jongens en meisjes?

Na interactie:
Dat was leuk hè? Vonden jullie Robin lief? Vinden jullie het leuk als Robin nog een keertje langskomt? Willen jullie nog iets vragen aan ons of aan Robin?

(Hier zou je kinderen de robot in kunnen laten stoppen met een dekentje als dat beschikbaar is, en dan afscheid laten nemen door de robot over z'n bolletje te laten aaien)

Appendix E – Anthropomorphism Questionnaire form (Dutch)

TILBURG UNIVERSITY *Datum Pre-/Post-test* / *ID-nummer*

Vragenlijst perceptie robot

Ik ga je nu een paar vragen stellen over Robin de robot. Ik ben benieuwd wat jij van Robin de robot denkt!

1. Denk je dat Robin de robot dingen kan zien?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

2. Denk je dat Robin de robot verdrietig kan zijn?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

3. Denk je dat Robin de robot iets kan onthouden?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

4. Denk je dat Robin de robot het voelt als je Robin kietelt?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

5. Denk je dat Robin pijn kan hebben?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

6. Vind je dat Robin de robot veel weet?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

7. Denk je dat Robin de robot het begrijpt als je iets zegt?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

8. Denk je dat Robin de robot groeit?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

TILBURG UNIVERSITY *Datum Pre-/Post-test* / *ID-nummer*

9. Denk je dat Robin de robot blij kan zijn?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

10. Vind je dat Robin de robot slim is?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

11. Denk je dat Robin de robot moet eten?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

12a. Zou je het leuk vinden om een spelletje te spelen met Robin?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test (Zie 12b)

12b. Zou je het leuk vinden om dit spelletje nog een keer met Robin te spelen?

Pre-test (zie 12a)

Post-test
Ja / Nee / Weet niet, omdat...

13. Zou je het leuk vinden om andere dingen te leren door met Robin spelletjes te spelen?

Pre-test
Ja / Nee / Weet niet, omdat...

Post-test
Ja / Nee / Weet niet, omdat...

Appendix G – Papercraft NAO Models

Robin de Robot

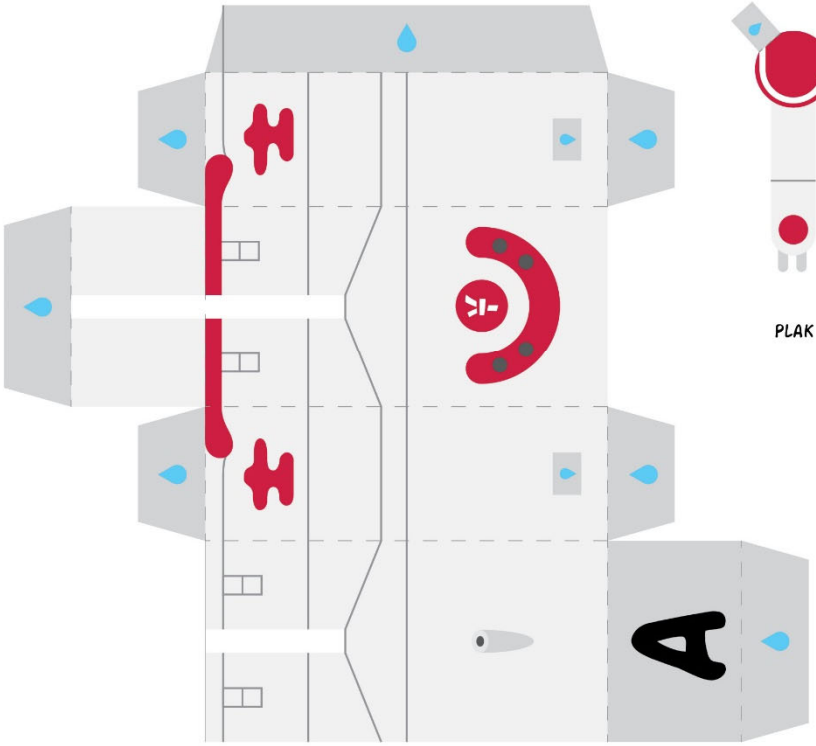
Hoi! Ken je mij nog? Mijn naam is Robin de robot! Kort geleden hebben wij met elkaar een spelletje gespeeld. Ik vond dat heel erg leuk en ik hoop dat jij dat ook vond. Als een bedankje wil ik je dit geven. Hiermee kun jij een papieren poppetje in de vorm van mij maken!

Dank je wel dat ik met jou een spelletje mocht spelen. Hopelijk zien wij elkaar nog een keer! Tot ziens!

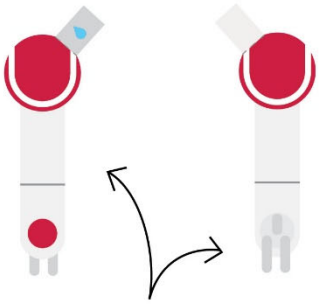
TIPS:

VOUWEN
LIJMEN (MET EEN LIJMSTIFT)
A A PLAK AAN ELKAAR

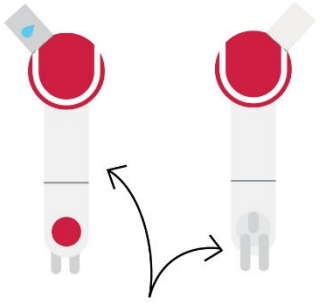
LICHAAM



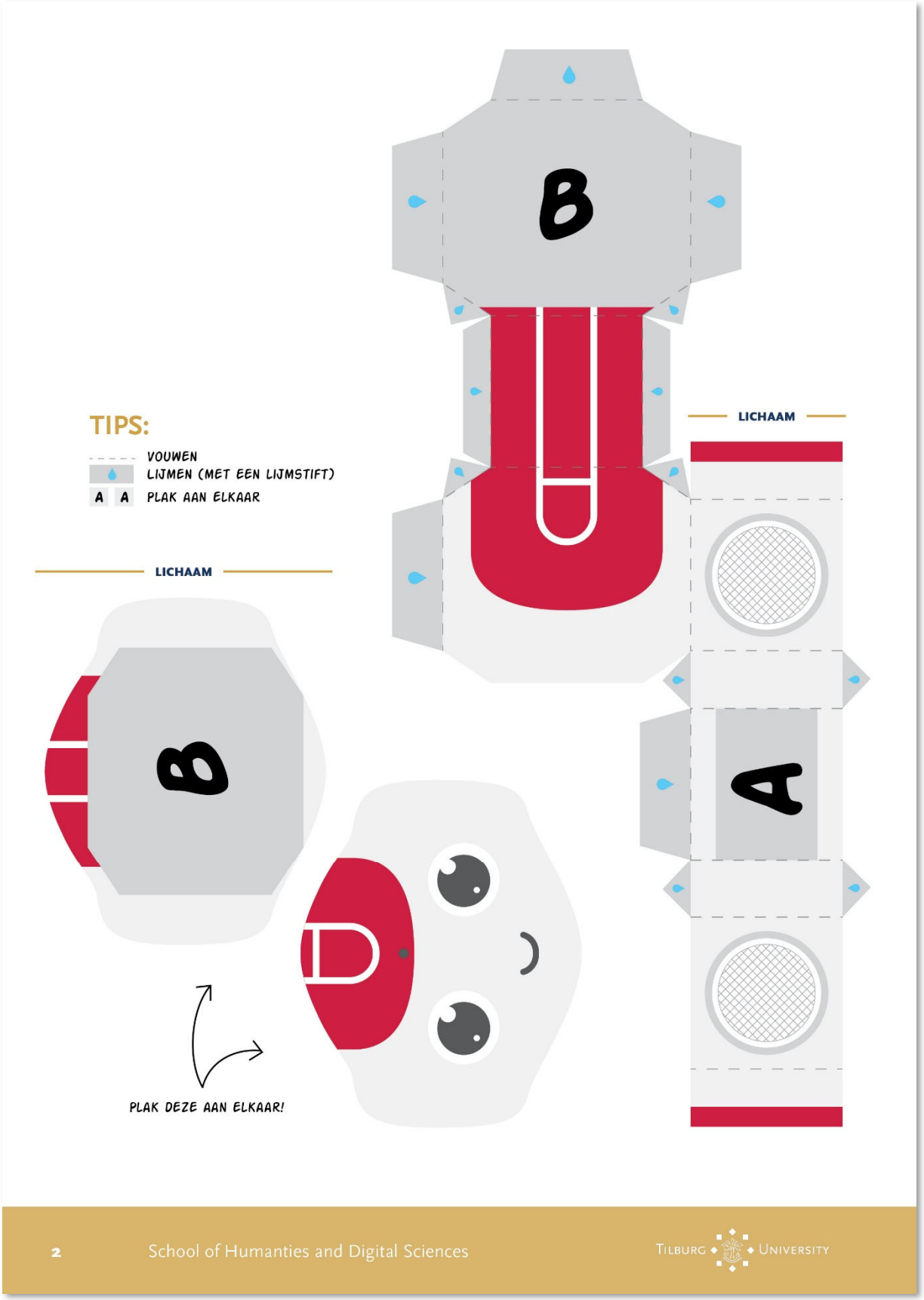
ARMEN



PLAK DEZE AAN ELKAAR!

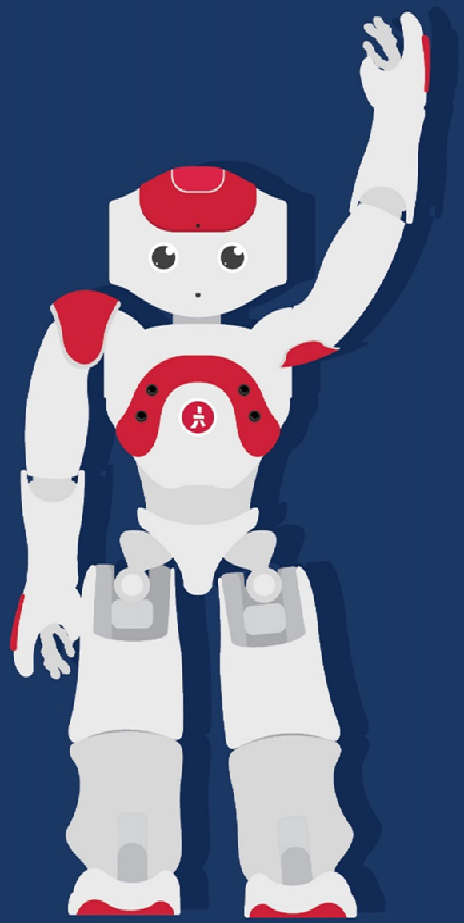


PLAK DEZE AAN ELKAAR!


TIPS:
VOUWEN
LIJMEN (MET EEN LIJMSTIFT)
A A PLAK AAN ELKAAR
LICHAAM
B
LICHAAM
B
A
PLAK DEZE AAN ELKAAR!
2
School of Humanties and Digital Sciences
Tilburg University

Appendix F – Infographic Sent to Schools and Parents (Dutch)

HET ONDERZOEK



Beste ouders of verzorgers! Misschien kent u mij nog, mijn naam is Robin de Robot. Kort geleden heb ik samen met uw kind een spel gespeeld om samen met Tilburg University te onderzoeken hoe sociale robots beter ingezet kunnen worden in het toekomstige onderwijs.

Destijds heeft u aangegeven dat u meer informatie over de uitkomsten van dit onderzoek wilde hebben, en daarom kom ik u daar vandaag iets over vertellen!

Voor dit onderzoek heb ik samen met uw kind een spelletje gespeeld. Samen met hen speelde ik, *"Ik zie, ik zie, wat jij niet ziet..!"*

Aan de hand van dit spelletje heb ik uw kind zes Engelse woorden proberen te leren.

Naast dat we hebben gekeken hoe goed uw kind het Engels oppakte in ons spel, hebben we ook gekeken hoe betrokken uw kind was met het spel, met mij en hoe actief uw kind hierin was.

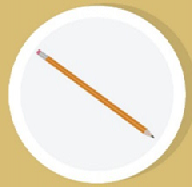
Als laatste hebben wij ook nog gekeken naar hoe menselijk uw kind mij vond. In andere woorden, zagen zij mij als een echt, levend wezen?

DE WOORDJES

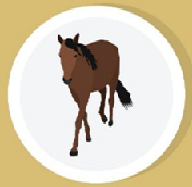
Brug / Bridge

Trap / Stairs

Lepel / Spoon



Potlood / Pencil

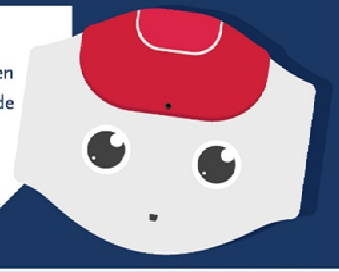


Paard / Horse

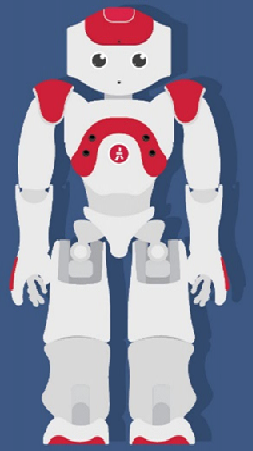
Schildpad / Turtle

DRIE VERSCHILLENDE GROEPEN

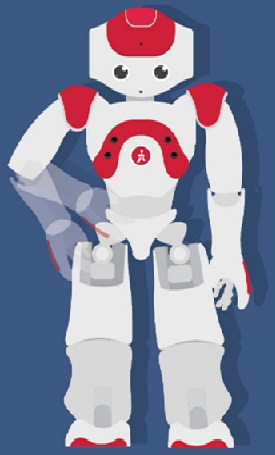
In het bijzonder wilden wij kijken of het gebruik van gebaren een robot kan helpen bij het beter aanleren van een tweede taal. Daarom werd uw kind willekeurig in één van drie groepen ingedeeld.



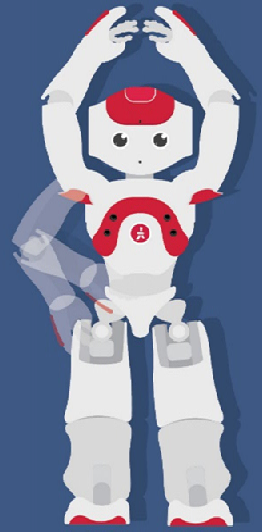
Ik zie, ik zie wat jij niet ziet en het is een...**POTLOOD**!



De 'Geen Gebaren' groep
In deze groep maakte ik geen gebaren, maar vroeg ik enkel om een ding dat ik zag.



De 'Enkel Gebaar' groep
In deze groep maakte ik één enkel gebaar per woord, welke ik elke keer dat het woord langskwam herhaalde.



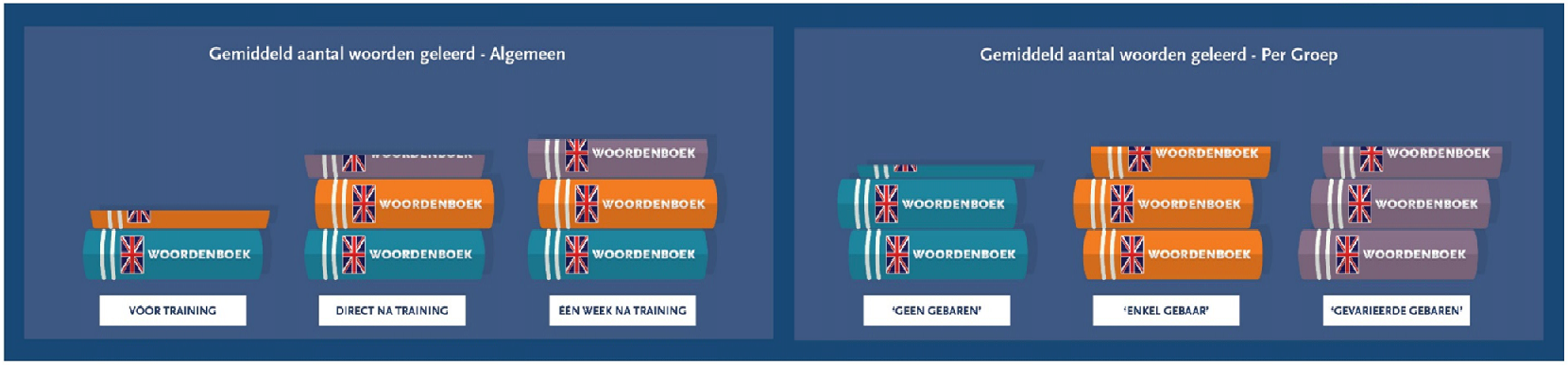
De 'Gevarieerde Gebaren' groep
In deze groep maakte ik voor elk woord een nieuw gebaar dat uw kind nog niet eerder gezien had.

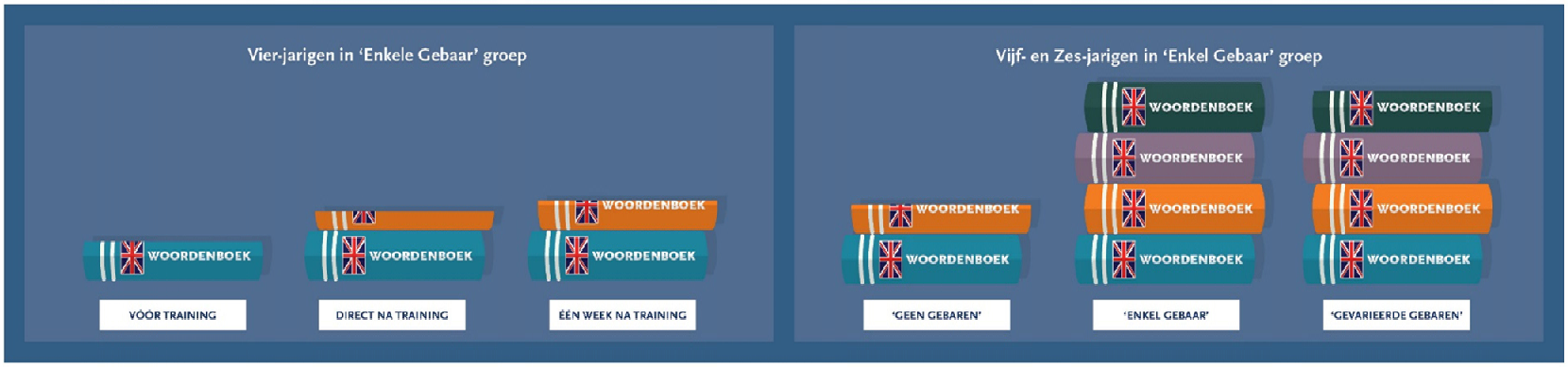
WOORDJES LEREN

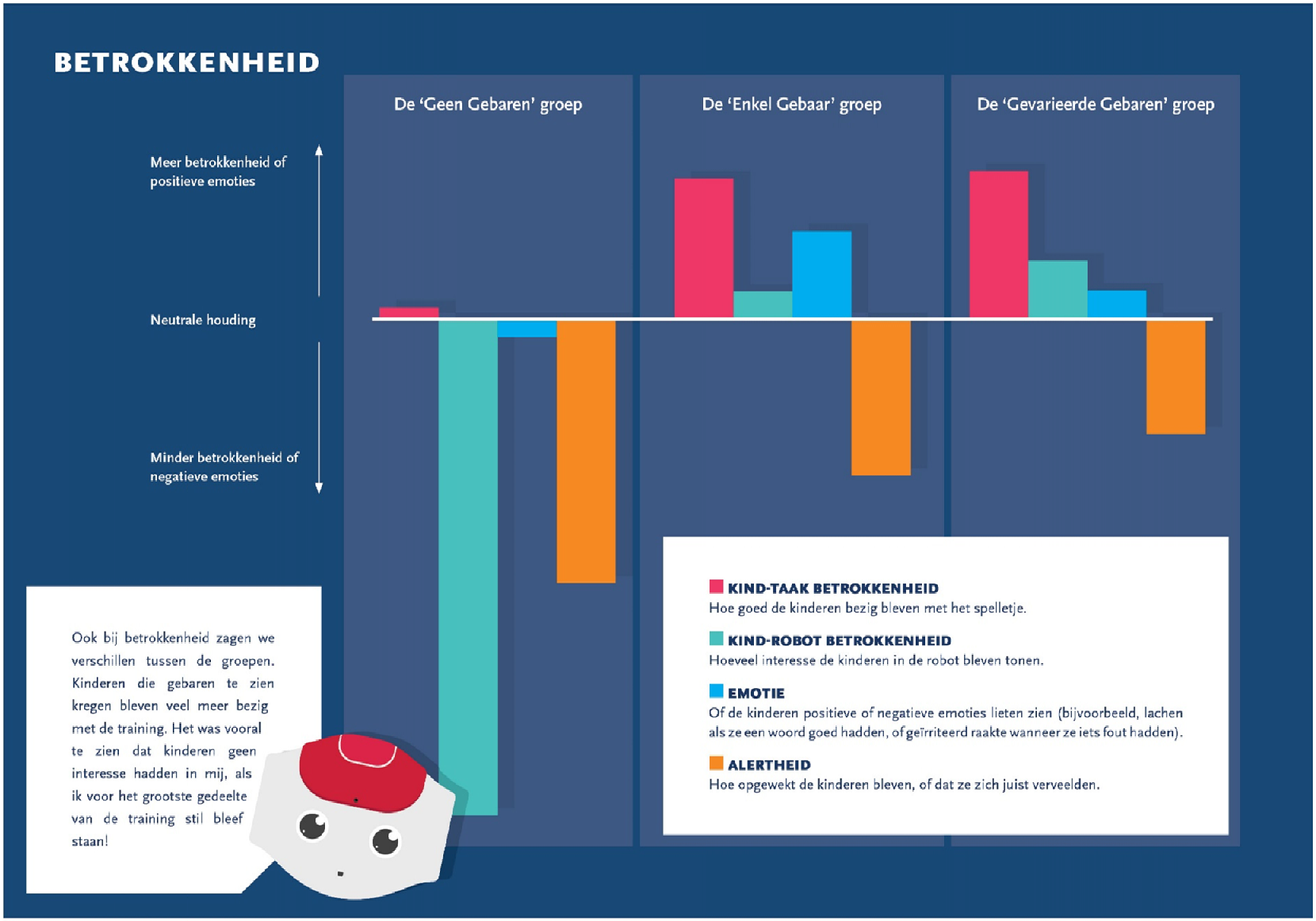
Als eerste hebben we gekeken naar hoeveel woorden alle kinderen hebben geleerd. Gelukkig hebben alle kinderen gemiddeld gezien iets geleerd! Maar gek genoeg zagen wij bijna geen verschil tussen de drie groepen, dat hadden wij juist wel verwacht...

Maar...! Als we wat specifieker gaan kijken naar de verschillende leeftijden, dan zien we ineens wel verschillen! Nu blijkt het ineens dat kinderen die vijf of zes zijn, veel meer baat hebben bij gebaren dan kinderen die vier jaar zijn. Dat is erg interessant!

Om een of andere reden werken gebaren voor hele jonge kinderen dus averechts.






BETROKKENHEID
De 'Geen Gebaren' groep
De 'Enkel Gebaar' groep
De 'Gevarieerde Gebaren' groep
Meer betrokkenheid of positieve emoties
Neutrale houding
Minder betrokkenheid of negatieve emoties
Ook bij betrokkenheid zagen we verschillen tussen de groepen. Kinderen die gebaren te zien kregen bleven veel meer bezig met de training. Het was vooral te zien dat kinderen geen interesse hadden in mij, als ik voor het grootste gedeelte van de training stil bleef staan!
KIND-TAAK BETROKKENHEID
Hoe goed de kinderen bezig bleven met het spelletje.
KIND-ROBOT BETROKKENHEID
Hoeveel interesse de kinderen in de robot bleven tonen.
EMOTIE
Of de kinderen positieve of negatieve emoties lieten zien (bijvoorbeeld, lachen als ze een woord goed hadden, of geïrriteerd raakte wanneer ze iets fout hadden).
ALERTHEID
Hoe opgewekt de kinderen bleven, of dat ze zich juist verveelden.

MENSELIJKHEID

'Biologische' vragen

Denk je dat Robin...	Percentage Ja-antwoord
...het voelt als je hem kietelt?	63 %
...pijn kan voelen?	49 %
...dingen kan zien?	83 %
...kan groeien?	34 %
...eten nodig heeft?	38 %

'Mentale' vragen

Denk je dat Robin...	Percentage Ja-antwoord
...het begrijpt als je iets tegen hem zegt?	84 %
...verdrietig kan zijn?	56 %
...iets kan onthouden?	66 %
...veel weet?	79 %
...slim is?	87 %
...blij kan zijn?	93 %

Bij de vragenlijst van hoe menselijk de kinderen mij vonden zagen wij gek genoeg geen verschillen tussen de groepen, of tussen de scores vóór en ná de training.

Wat we wel zagen was meeste kinderen bij de 'mentale' vragen de antwoorden meer op zichzelf betrokken ("Robin is slim, want hij gaat ook naar school!"), terwijl ze bij de biologische vragen meer lette op wat ze zagen ("Robin kan tegen kietelen, want hij is van metaal!").

Daarnaast wilden bijna alle kinderen nog een keer opnieuw met mij spelen, dus daar word ik erg blij van!

CONCLUSIE

Alles bij elkaar hebben wij erg veel geleerd van dit onderzoek. We vroegen ons af of robot gebaren kunnen helpen bij het leren, en ze lijken ook zeker iets te doen!

Wat we alleen niet helemaal hadden verwacht is dat oudere kinderen hier veel meer baat bij hadden dan jonge kinderen. Natuurlijk zullen oudere kinderen iets sneller leren dan jonge kinderen, maar dat het verschil zo groot zou zijn is onverwacht.

Misschien dat het zien van mij als robot zoveel indruk heeft gemaakt bij de jonge kinderen, dat ze gewoon vergeten waren om de woordjes te onthouden!

In ieder geval geeft dit weer nieuwe stof tot nadenken. En zal ik naar nog meer scholen toe moeten gaan, om er achter te komen of dit verschil nog vaker voor komt.

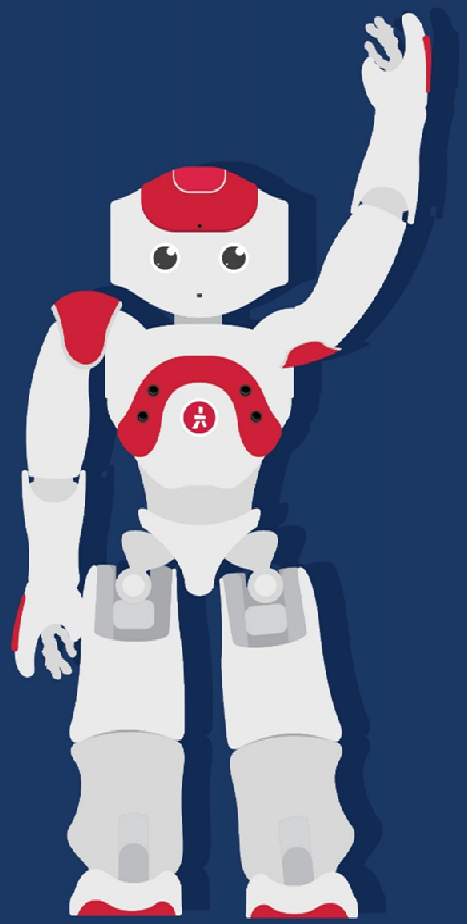
Namens Tilburg universiteit, en ook namens mij, willen wij beide scholen die ons hebben geholpen enorm bedanken. Zonder Nutsbasisschool Teteringen en Basisschool de Stappen in Tilburg hadden wij dit onderzoek nooit kunnen doen.

Uiteraard willen wij ook alle ouders en verzorgers bedanken voor hun interesse en enthousiasme om mee te doen aan dit onderzoek.

En als laatste willen wij ook alle kinderen heel erg bedanken dat zij samen met mij Engels wilden leren. Ik vond het enorm leuk, en ik hoop dat zij dat ook vonden!

Voor nu ga ik weer terug naar de universiteit, maar wie weet zien wij elkaar nog een keer in een van de volgende schooljaren!

Groetjes,
Robin de Robot



Mocht u interesse hebben in meer onderzoeken samen met Robin de Robot, dan verwijzen wij u graag naar de websites van L2TOR, *www.l2tor.eu*

Uiteraard kunt altijd contact met ons opnemen als u vragen of opmerkingen heeft. U kunt hiervoor terecht bij project-leider Jan de Wit, *j.m.s.dewit@tilburguniversity.edu*

Appendix H – Descriptive Statistics for Word Learning

Table H1. Vocabulary Learning scores per condition, *M (SD)*

Factors	*No-Gestures*	*Single-Gesture*	*Varied-Gestures*	**Total**
Pre-test	1.15 (1.00)	1.19 (1.06)	1.66 (1.14)	**1.32 (1.08)**
Immediate Post-test	2.21 (1.85)	2.78 (2.12)	2.31 (1.61)	**2.44 (1.88)**
Delayed Post-test	2.64 (1.50)	2.78 (1.93)	2.79 (1.63)	**2.73 (1.68)**

Note. Chance level was calculated at 0.44 words.

Appendix I – Elaborate Results for Anthropomorphism

Table I1 features the questions of the anthropomorphism questionnaire, accompanied by the means and standard deviations per question. In general, it can be said that the children anthropomorphised the robot on the questionnaire in both the pre- and post-test measuring moments (Figure I1a), scoring a 7.32 (*SD* = 2.49) and 7.32 (*SD* = 2.71) respectively on a combined scale of 0 – 11. Each individual question was scored either as a 0 (a 'no' or 'I don't know' answer) or a 1 (a 'yes' answer), with any mean score above .50 showing that on average, more children anthropomorphised the robot on that particular question.

Table I1. Anthropomorphism Questionnaire Scores Pre- and Post-test, *M (SD)*

	Do you think that Robin…	*Pre-test*	*Post-test*
Biological			
1.	…can feel it if you tickle him?	.63 (.49)	.53 (.50)
2.	…can feel pain?	.49 (.50)	.41 (.50)
3.	…can see things?	.83 (.38)	.84 (.37)
4.	…grows?	.34 (.48)	.41 (.50)
5.	…needs food?	.38 (.49)	.40 (.49)
Mental			
6.	…understands it when you say something?	.84 (.37)	.80 (.40)
7.	…can be sad?	.56 (.50)	.52 (.50)
8.	…can remember something?	.66 (.48)	.72 (.45)
9.	…knows a lot?	.79 (.41)	.86 (.35)
10.	…is smart?	.87 (.34)	.90 (.30)
11.	…can be happy?	.93 (.26)	.90 (.30)
Other			
12.	Would you like to play 'I spy with my little eye' again with Robin?	1.00 (.00)	.94 (.25)
13.	Would you like to learn other things by playing games with Robin?	.95 (.23)	.99 (.10)
Total scores			
	Biological*	2.67 (1.56)	2.61 (1.65)
	Mental**	4.65 (1.36)	4.71 (1.41)
	Overall***	7.32 (2.49)	7.32 (2.71)

Notes. * Score out of 5; ** Score out of 6; *** Score out of 11, "Other" questions are excluded

In general, children anthropomorphized highly on the questions 'can see things', 'understands when you say something', 'knows a lot', 'is smart', 'can be happy', scoring higher than .80 on these questions pre-test. Children anthropomorphised low on the 'grows' and 'needs food' questions, scoring lower than .40 in the pre-test. Furthermore, children had more difficulties with anthropomorphising biological traits (scoring 2.67 (*SD* = 1.56) and 2.61 (*SD* = 1.65) on pre- and post-test, with a maximum score of 5) than with mental traits (scoring 4.65 (*SD* = 1.36) and 4.71 (*SD* = 1.41) on pre- and post-test, with a maximum score of 6). Finally, all children indicated they were excited to play a game with Robin during the pre-test (*M* = 1.00, *SD* = .00), and mostly wanted to play the game again post-experiment (*M* = .94, *SD* = .25). Similarly, nearly all children stated they would also enjoy playing different types of learning games with Robin (*M* = .95, *SD* = .23) pre-test, and even more so after the post-test (*M* = .99, *SD* = .10). The relatively high standard deviations also show the large variations between individual children (as also visualised by Figure I1b).

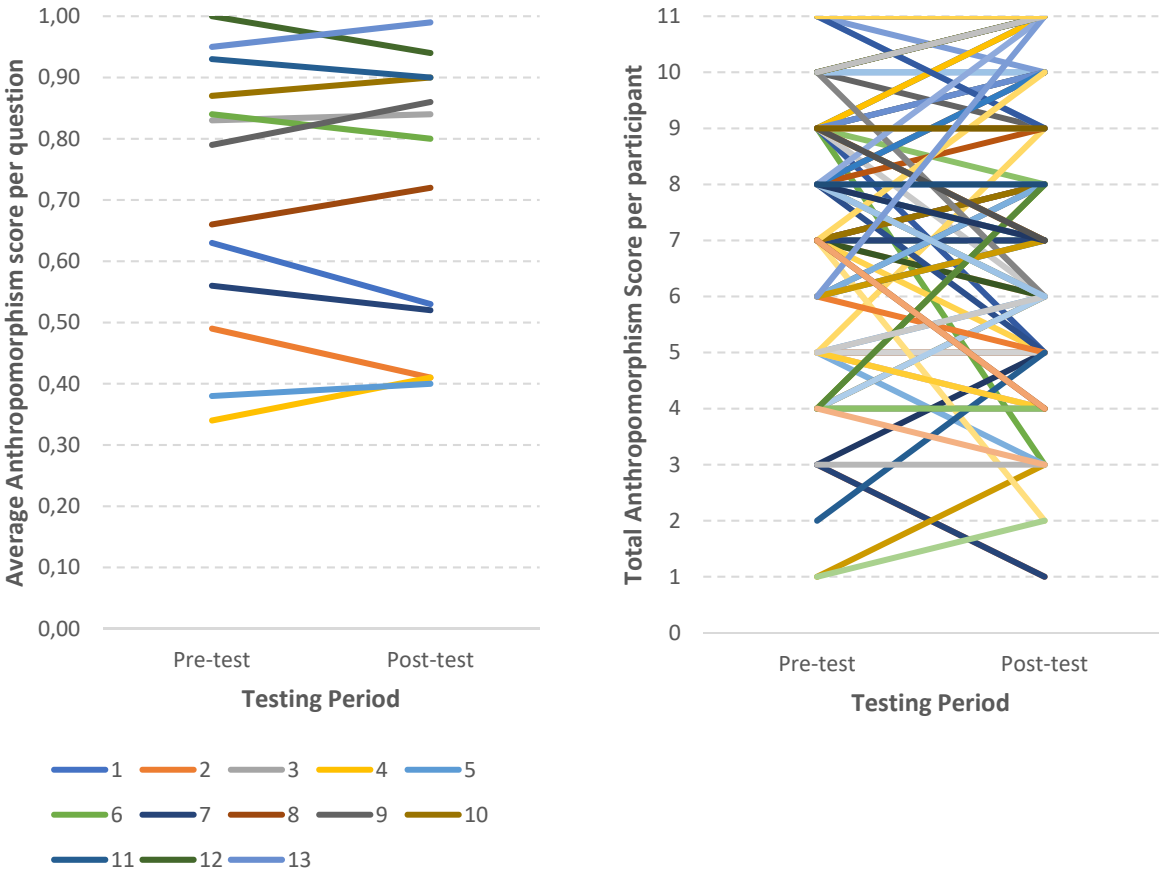


Figure I1. Figures showing overall variability per question (H1a) and individual variance per participant (H1b) between pre- and post- anthropomorphism test.

An analysis using a Pearson's Correlation on whether the children's perception of the robot changed showed a moderately large correlation, $r = .71$, $p < .001$. 50.1% of the variance in post-test scores was accounted for by the scores of the pre-test (Figure I2). In other words, children who anthropomorphised the robot pre-experiment, generally did so too post-experiment. Overall, most children were fairly consistent to what degree they anthropomorphised the robot, though some children changed their opinion substantially between pre- and post-tests, as can be seen in Figure I3.

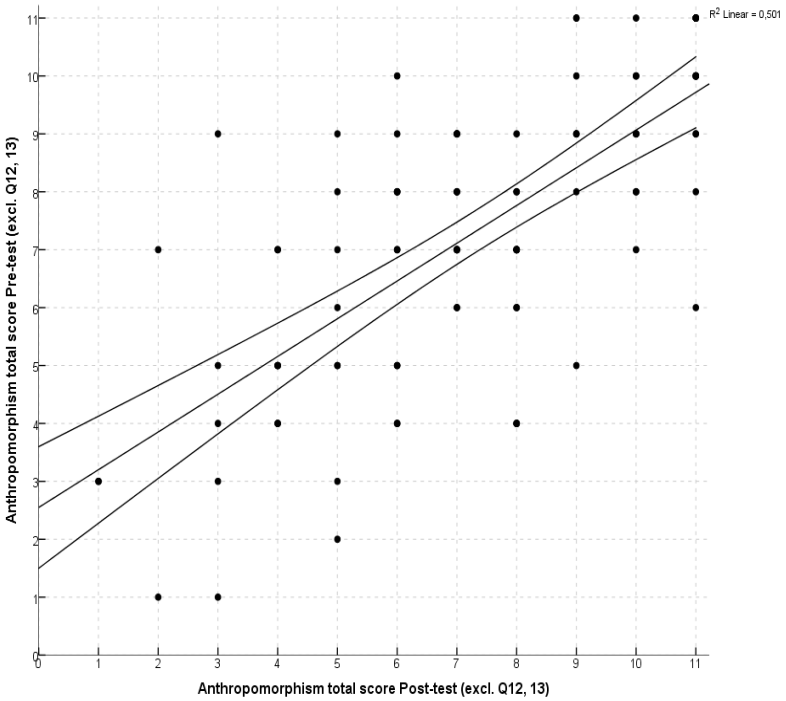


Figure I2. Graph showing correlation between pre- and post-test anthropomorphism scores.

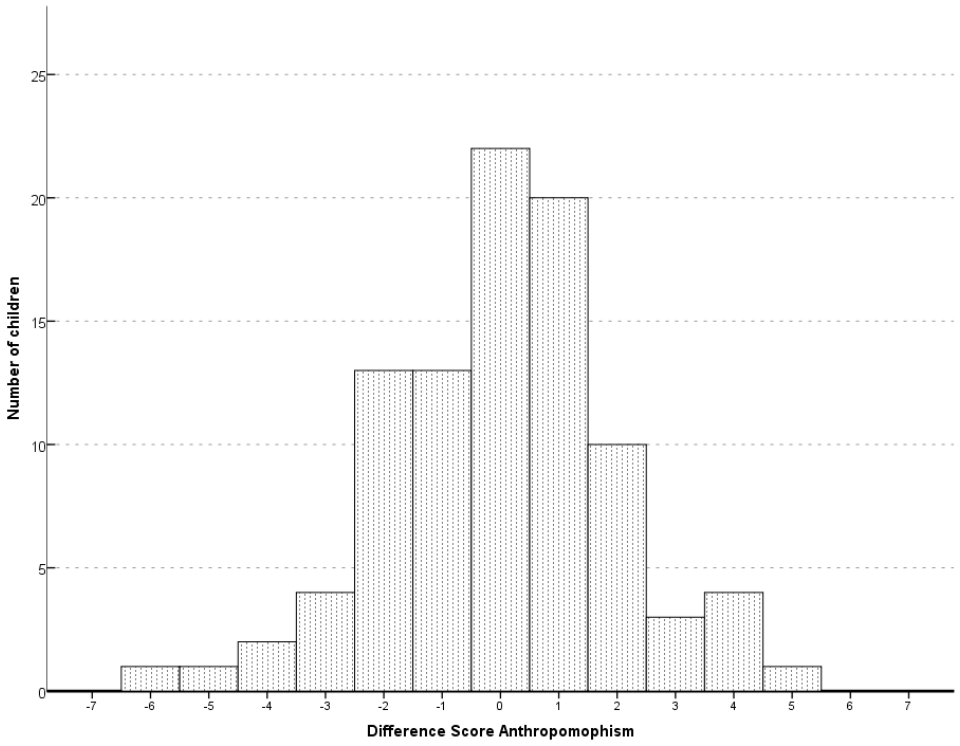


Figure I3. Histogram showing overall difference for each participant's pre- and post- anthropomorphism questionnaire scores.

Appendix J – Elaborate Results for Engagement per Age Group

Average Engagement per age group

First of all, the average engagement scores over the entire training were reanalysed based on the new age groups, by performing multiple factorial ANOVAs per engagement construct. When looking at child-task engagement, no main effect was found for the two age groups $F(1, 88) = 1.18$, $p = .281$. Similarly, no interaction effect was found between age and condition, $F(2, 88) = 0.67$, $p = .512$. For child-robot engagement a similar outcome was found: there was no main effect for the age groups, $F(1, 88) = 0.19$, $p = .664$ nor was there an interaction effect, $F(2, 88) = 0.23$, $p < .664$. Arousal too found no main effect ($F(1, 88) = 0.19$, $p = .666$), nor an interaction effect ($F(2, 88) = 0.97$, $p = .383$. Finally, valence showed no main effect for the age groups, $F(1, 88) = 0.55$, $p = .459$. However, valence did show a medium-sized interaction effect between condition and age, $F(2, 88) = 5.93$, $p = .004$, $\eta_p^2 = .119$. A simple effects analysis showed that there was a difference in average valence score between age groups in the varied-gestures condition ($M_{dif} = 0.34$, $p = .003$). On average, children in age group five and six showed more negative emotions ($M = -0.07$, $SD = 0.20$) in the varied-gestures condition than those aged four ($M = 0.28$, $SD = 0.30$).

Engagement decline between rounds per age group

A second analysis was also done to look more specifically at the decline in engagement between the two rounds for each age-group and condition. To examine this, several repeated-measures ANOVAs were performed. As these results are fairly extensive, only the main- and other notable points will be discussed here. Descriptive statistics (Table J2) and all ANOVA outputs (Table J3) can be found in in the tables

For both age groups, main effects were found for all constructs ($p < .010$), overall the effect sizes were very large ($\eta_p^2 > .180$ for valence, $\eta_p^2 > .580$ for all other constructs). No interaction effects were found. As can be seen in Table J1, similar to the non-segmented analysis in section 4.2.2, each engagement construct declined as the training went on for both age groups.

Concludingly it can be said that gestures led to higher child-robot

Table J1. Decline of engagement scores between round 4 and round 24. Pairwise comparison of Main Effect between rounds for each construct within each age group.

Factors	M_{dif}	*p*
Age 4		
Child-Task (between rounds)	-0.71	< .001
Child-Robot (between rounds)	-0.57	< .001
Arousal (between rounds)	-0.68	< .001
Valence (between rounds)	-0.23	.007
Ages 5 and 6		
Child-Task (between rounds)	-0.62	< .001
Child-Robot (between rounds)	-0.56	< .001
Arousal (between rounds)	-0.57	< .001
Valence (between rounds)	-0.16	.001

engagement over time for both age groups. In terms of arousal, children aged four were more aroused over the course of the training when presented with gesture variations. And finally, children aged five and six exhibited more positive emotions over the course of the entire training when presented with a robot using repeated gestures, compared to a robot using no or variation in gestures.

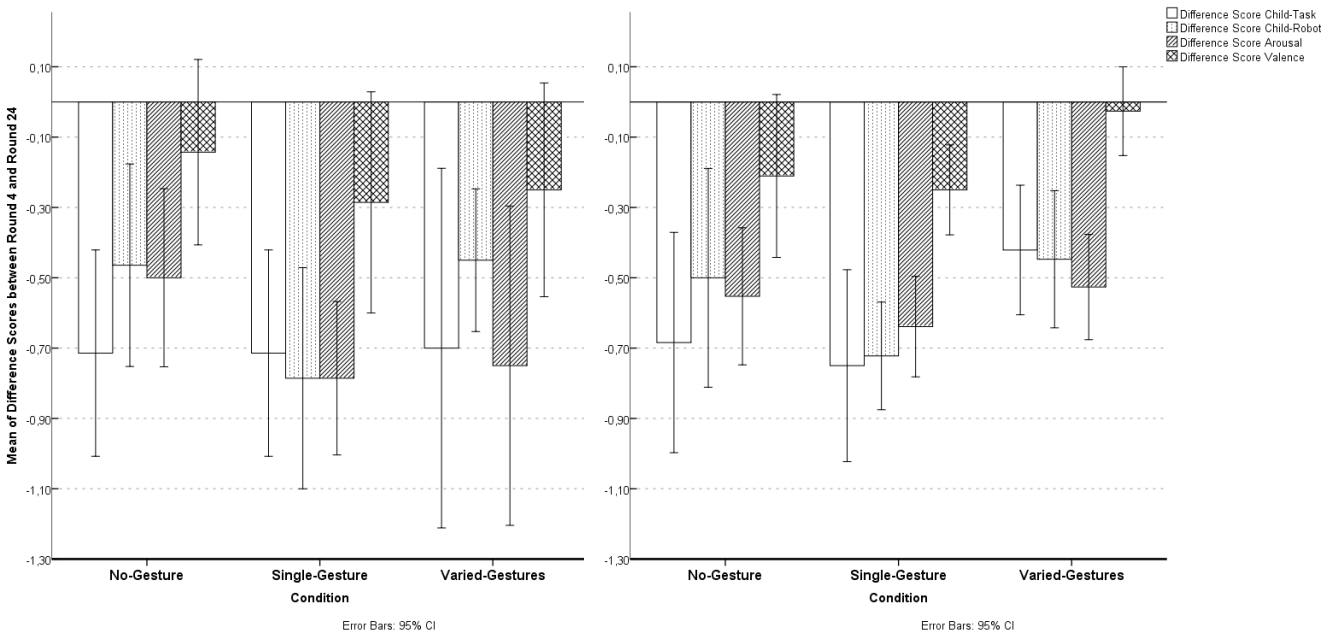


Figure J1. Difference scores of separate engagement constructs between conditions in age groups (a) four, and (b) five and six.

Table J2. Engagement scores per age group and condition, *M (SD)*

Factors	*No-Gesture*			*Single-Gesture*			*Varied-Gestures*		
	Round 4	Round 24	**Average**	Round 4	Round 24	**Average**	Round 4	Round 24	**Average**
Age 4									
Child-Task	0.21 (0.58)	-0.50 (0.71)	**-0.14 (0.59)**	0.50 (0.65)	-0.21 (0.38)	**0.14 (0.70)**	0.65 (0.38)	-0.05 (0.72)	**0.30 (0.44)**
Child-Robot	-0.61 (0.76)	-1.07 (0.55)	**-0.84 (0.62)**	0.54 (0.66)	-0.25 (0.87)	**0.14 (0.73)**	0.30 (0.54)	-0.15 (0.41)	**0.08 (0.46)**
Arousal	-0.25 (0.48)	-0.75 (0.38)	**-0.50 (0.34)**	0.11 (0.53)	-0.68 (0.37)	**-0.29 (0.41)**	0.30 (0.59)	-0.45 (0.28)	**-0.08 (0.33)**
Valence	0.00 (0.39)	-0.14 (0.36)	**-0.07 (0.30)**	0.21 (0.38)	-0.07 (0.58)	**0.07 (0.41)**	0.40 (0.46)	0.15 (0.24)	**0.28 (0.30)**
Ages 5 and 6									
Child-Task	0.47 (0.56)	-0.21 (0.69)	**0.13 (0.54)**	0.69 (0.42)	-0.06 (0.64)	**0.32 (0.47)**	0.45 (0.55)	0.03 (0.68)	**0.24 (0.59)**
Child-Robot	-0.63 (0.78)	-1.13 (0.64)	**-0.88 (0.64)**	0.33 (0.62)	-0.39 (0.65)	**-0.03 (0.62)**	0.34 (0.44)	-0.11 (0.70)	**0.12 (0.55)**
Arousal	-0.16 (0.37)	-0.71 (0.42)	**-0.43 (0.34)**	0.06 (0.42)	-0.58 (0.43)	**-0.26 (0.40)**	0.00 (0.37)	-0.53 (0.39)	**-0.26 (0.35)**
Valence	0.11 (0.36)	-0.11 (0.21)	**0.00 (0.17)**	0.33 (0.34)	0.08 (0.31)	**0.21 (0.30)**	-0.05 (0.28)	-0.08 (0.19)	**-0.07 (0.20)**

Table J3. ANOVA outputs for all measured engagement outputs within each age group.

Factors	*df*	*F*	*p*	η_p^2
Age 4				
Child-Task (between rounds)	1, 35	57.71	< .001 [I4]	.622
Child-Task (between rounds and conditions)	2, 35	0.00	.998 [I4]	-
Child-Robot (between rounds)	1, 35	53.30	< .001 [I5]	.604
Child-Robot (between rounds and conditions)	2, 35	2.13	.135*[I5]	-
Arousal (between rounds)	1, 35	74.82	< .001 [I6]	.681
Arousal (between rounds and conditions)	2, 35	1.44	.251*[I6]	-
Valence (between rounds)	1, 35	8.10	.007 [I7]	.188
Valence (between rounds and conditions)	2, 35	0.33	.725 [i7]	-
Ages 5 and 6				
Child-Task (between rounds)	1, 53	73.84	< .001 [I4]	.582
Child-Task (between rounds and conditions)	2, 53	1.96	.152 [I4]	-
Child-Robot (between rounds)	1, 53	76.19	< .001*[I5]	.590
Child-Robot (between rounds and conditions)	2, 53	1.72	.190 [I5]	-
Arousal (between rounds)	1, 53	159.77	< .001*[I6]	.751
Arousal (between rounds and conditions)	2, 53	0.55	.578 [I6]	-
Valence (between rounds)	1,53	11.98	.001 [I7]	.184
Valence (between rounds and conditions)	2, 53	2.17	.124*[I7]	-

Notes. Letters after *p*-value denote specific table below with pairwise comparisons and / or Post-Hoc tests. Asterisks denotes a non-significant overall value, that did show a significant result in a particular group pairing in a Post Hoc analysis.

Table J4. Child Task Engagement. Pairwise comparison of Main Effect between rounds. Post-Hoc Bonferroni between rounds and conditions.

Factors	M_{dif}	*p*
Age 4		
Child-Task (between rounds)	-0.71	< .001
Between NG and SG	-0.29	.651
Between NG and VG	-0.44	.252
Between SG and VG	-0.16	1.000
Ages 5 and 6		
Child-Task (between rounds)	-0.62	< .001
Between NG and SG	-0.19	.873
Between NG and VG	-0.11	1.000
Between SG and VG	0.08	1.000

Table J5. Child Robot Engagement. Pairwise comparison of Main Effect between rounds. Post-Hoc Bonferroni between rounds and conditions.

Factors	M_{dif}	*p*
Age 4		
Child-Robot (between rounds)	-0.57	< .001
Between NG and SG	-0.98	.001
Between NG and VG	-0.91	.004
Between SG and VG	0.07	1.000
Ages 5 and 6		
Child-Robot (between rounds)	-0.56	< .001
Between NG and SG	-0.85	< .001
Between NG and VG	-1.00	< .001
Between SG and VG	-0.15	1.000

Table J6. Arousal. Pairwise comparison of Main Effect between rounds. Post-Hoc Bonferroni between rounds and conditions.

Factors	M_{dif}	*p*
Age 4		
Arousal (between rounds)	-0.68	< .001
Between NG and SG	-0.21	.397
Between NG and VG	-0.43	.025
Between SG and VG	-0.21	.526
Ages 5 and 6		
Arousal (between rounds)	-0.57	< .001
Between NG and SG	-0.17	.478
Between NG and VG	-0.17	.456
Between SG and VG	0.00	1.000

Table J7. Valence. Pairwise comparison of Main Effect between rounds. Post-Hoc Bonferroni between rounds and conditions.

Factors	M_{dif}	*p*
Age 4		
Valence (between rounds)	-0.23	.007
Between NG and SG	-0.14	.842
Between NG and VG	-0.35	.062
Between SG and VG	-0.20	.489
Ages 5 and 6		
Valence (between rounds)	-0.16	.001
Between NG and SG	-0.21	.023
Between NG and VG	0.07	1.000
Between SG and VG	0.27	.002

Notes. NG = No-Gestures condition. SG = Single-Gesture Condition. VG = Varied-Gestures condition.

Appendix K – Elaborate Results for Anthropomorphism per Age Group

When taking age into consideration, a repeated-measures ANOVA found that children aged four showed no significant difference in pre-test scores (M = 7.95, SD = 2.51) and post-test scores (M = 7.53, SD = 2.91), $F(1, 35) = 2.11$, $p = .156$. Children aged five and six also showed no significant differences in pre-test scores (M = 6.89, SD = 2.40) and post-test scores (M = 7.18, SD = 2.57), $F(1, 53) = 0.98$, $p = .327$. No differences were found between the two groups either, $F(1, 92) = 2.90$, $p = .092$.

When again taking conditions into consideration, children aged four showed no significant differences between conditions, $F(2, 35) = 0.01$, $p = .982$, nor did the five- and six-year olds, $F(2, 53) = 0.18$, $p = .835$. Thus, age had no effect on how the children anthropomorphised the robot. For an overview of the means per age group, view Table K1.

Table K1. Perceived Anthropomorphism score per condition and age group, *M (SD)*

Factors	*No-Gestures*		*Single-Gesture*		*Varied-Gestures*	
	Pre-test	Post-test	Pre-test	Post-test	Pre-test	Post-test
Age 4						
Score	7.71 (2.87)	7.29 (3.69)	8.43 (1.95)	8.07 (2.17)	7.60 (2.84)	7.10 (2.77)
Ages 5 and 6						
Score	7.00 (2.81)	7.16 (2.48)	6.50 (1.95)	6.67 (2.28)	7.16 (2.43)	7.68 (2.95)

Note. Scale ranges from 0 to 11.

Appendix L – Elaborate Analysis of Experiment Duration and Error-rate

For this analysis, normality was violated (Duration $z_{skewness}$ = 3.56; Error-rate $z_{skewness}$ = 2.01). On average, children took 16.89 minutes (*SD* = 4.07 minutes) to follow the training (excluding the initial practice round). When taking conditions into account, those in the no-gesture condition (M = 14.31 minutes, SD = 2.43 minutes) took less time to complete the training than those in the single-gesture (M = 17.80 minutes, SD = 4.48 minutes) and varied-gesture (*M* = 18.82 minutes, *SD* = 3.65 minutes) conditions, a two-way ANOVA showed that this difference was significant ($F(2, 88) = 18.30$, $p < .001$, $\eta_p^2 = .294$), representing a large-sized effect. The boxplot in Figure L1 shows the large variances in training durations between conditions. Similarly, a large-sized main effect was found for age ($F(1, 88) = 15.71$, $p < .001$, $\eta_p^2 = .151$), children who were in the four-year old age group (*M* = 21.33 minutes, *SD* = 4.78 minutes) took significantly longer to complete the training than those in the five- and six-year old age group (*M* = 18.30 minutes, *SD* = 3.40 minutes).

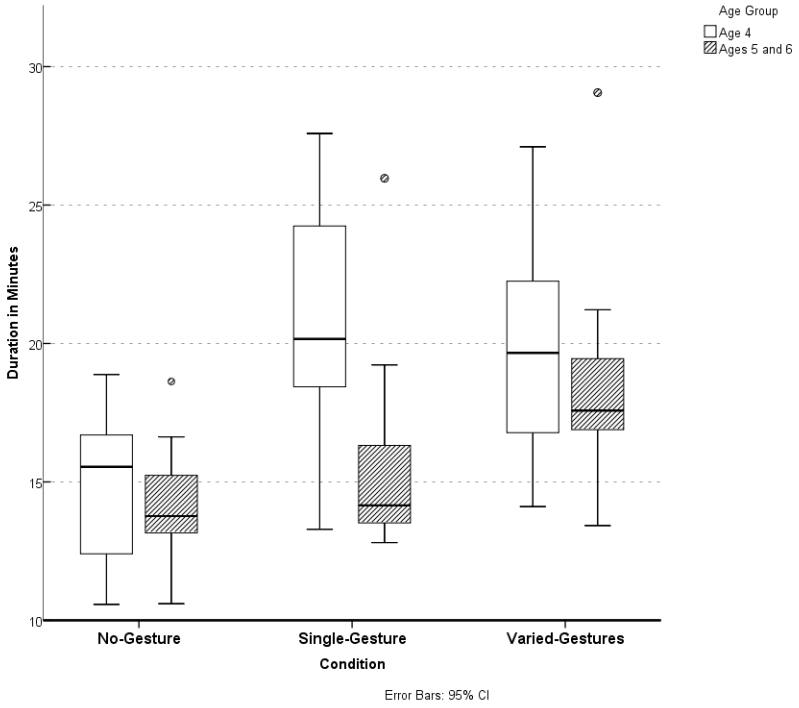


Figure L1. Boxplot showing overall duration per condition and age group.

Finally, a medium-sized interaction effect was found between age and condition ($F(2, 88) = 4.75$, $p = .011$, $\eta_p^2 = .097$). When looking at the differences between age groups within conditions, a simple effects analysis (visualised in Figure L2a) revealed a significant effect in the single-gesture condition between the two age groups, $F(1, 85) = 23.06$, $p < .001$. Children who were younger took significantly longer to complete the training in the single-gesture condition than the five- and six-year old children (M_{dif} = 5.52 minutes, *p*

< .001) No effects were found in the no-gesture condition ($F(1, 85) = 0.39$, $p = .534$) or in the varied-gestures condition ($F(1, 85) = 2.26$, $p = .136$).

When looking at the differences between conditions within the age groups, a simple effects analysis (visualised in Figure L2b) revealed a significant difference for both the four-year old children ($F(1, 85) = 14.63$, p < .001) and the five- and six-year old children ($F(1, 85) = 8.16$, p = .001). For the four-year old children, the no-gesture condition took significantly shorter to complete than both the single-gesture condition ($M_{dif} = 6.19$ minutes, $p < .001$) as well as the varied-gestures condition ($M_{dif} = 5.34$ minutes, $p < .001$). No difference was found between the single- and varied-gestures conditions ($M_{dif} = 0.85$ minutes, $p = .529$). For the five- and six-year old children, the no-gesture condition took significantly shorter than the varied-gesture condition ($M_{dif} = 4.15$ minutes, $p < .001$). Similarly, a significant result showed that the single-gesture condition took longer than the varied-gestures condition ($M_{dif} = 2.78$ minutes, $p = .010$). No difference was found between the no-gesture and the single-gesture condition ($M_{dif} = 1.37$ minutes, $p = .200$).

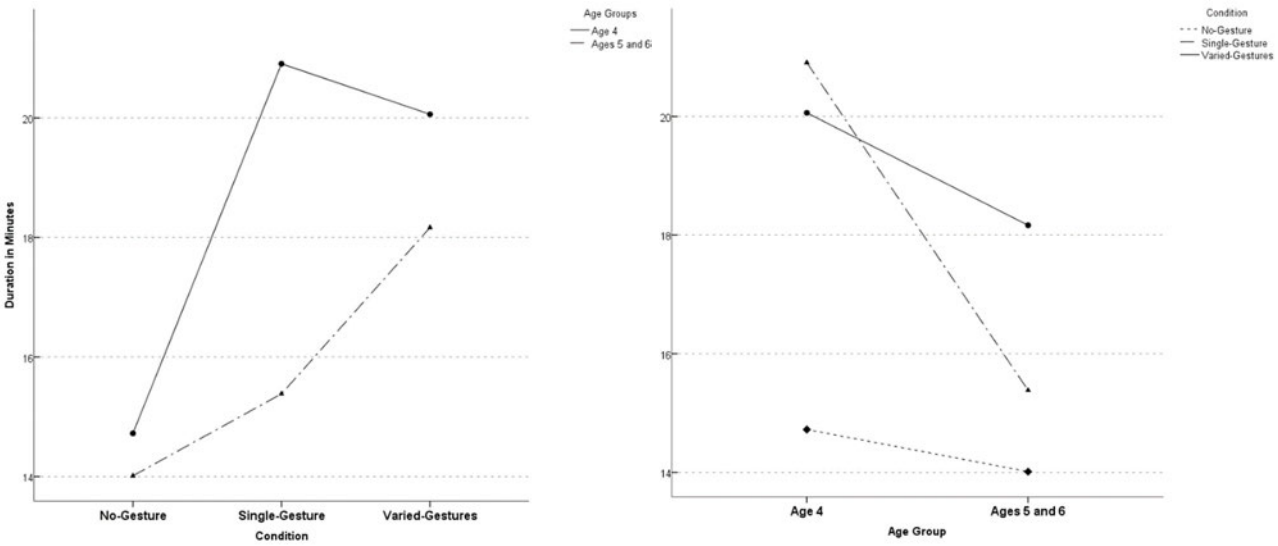


Figure L2. Line graph showing (L2a) the interaction effect in conditions between age groups, and (L2b) the interaction effect in the age groups between the three conditions.

To examine whether training completion times had an effect on immediate post-test scores, A Pearson correlation was performed, which showed a small-sized significant correlation, r = -.475, p < .001, 95% CI [-.591, -.346]. 22.6% of the variance in immediate post-test scores was accounted for by the time it took to complete the training. It can thus be concluded that those who took longer to complete the training were more likely to score lower on the post-test. Similar results were found when taking age into account; age four: r = -.388, p = .005, $r^2 = .151$, 95% CI [-.622, -.076], ages five and six: r = -.472, p < .001, $r^2 = .222$, 95% CI [-.613, -.305].

Table L1. Error rate per condition and age group, *M (SD)*

Factors	*No-Gestures*	*Single-Gesture*	*Varied-Gestures*	**Total**
Age 4				
Number of errors	15.86 (9.03)	15.79 (6.04)	14.50 (9.17)	**15.47 (7.89)**
Ages 5 and 6				
Number of errors	13.00 (5.84)	5.67 (7.04)	10.05 (6.07)	**9.64 (6.90)**
Total				
Number of errors	**14.21 (7.37)**	**10.09 (8.28)**	**11.59 (7.44)**	**12.00 (7.83)**

Finally, the means show a slight difference in number of errors made between the three conditions and the two age groups (Table L1), a two-way ANOVA revealed that there was no main effect for the number errors that were made between the three conditions, $F(2, 88) = 2.20$, $p = .117$. However, a large-sized main effect was found between the two age groups, $F(1, 88) = 15.01$, $p < .001$, $\eta_p^2 = .146$. Younger children made significantly more errors ($M = 15.47$, $SD = 7.89$) than older children ($M = 9.64$, $SD = 6.90$). No interaction effect was found between the age groups and the conditions, $F(2, 88) = 2.28$, $p = .108$. A simple effects analysis did however reveal a significant effect in the single-gesture condition between the two age groups. Children aged four made significantly more errors ($M_{dif} = 10.12$ errors, $p < .001$) in the single-gesture condition than those aged five and six. A second simple effects analysis also revealed that there was a significant difference in number of errors made in the five and six-year olds between the no-gesture condition and single-gesture condition, ($M_{dif} = 7.33$ errors, $p = .002$). This indicates that five- and six-year olds made significantly fewer errors when presented with repeated gestures, than younger children.

Finally, a significant correlation was also found between number of errors made and training duration, $r = .72$, $p < .001$, 95% CI [.62, .80]. 51.8% of the variances in duration was explained for by the error rate. Furthermore, as can be seen in Figure L3, the split between the two scatter-groups is mostly caused by the lower training duration in the no-gesture condition.

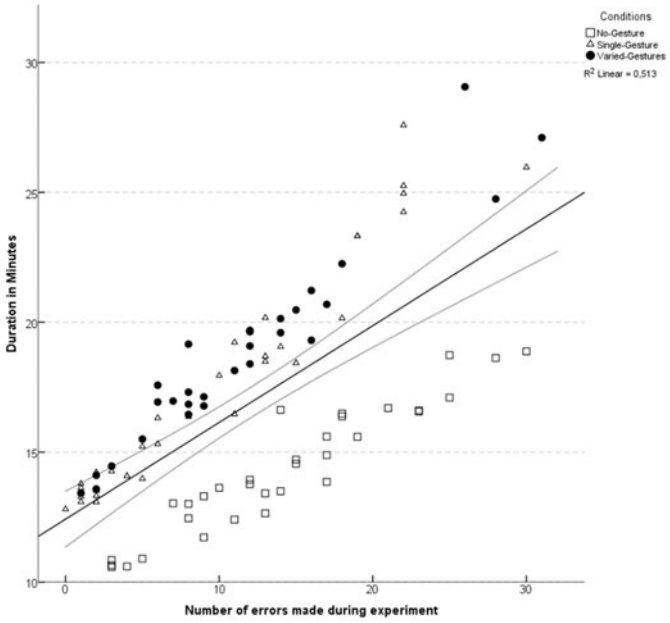


Figure L3. Scatterplot showing correlation between number of errors and duration per condition.