
# THE EFFECT OF EMBODIMENT ON ADAPTIVE ROBOT ASSISTED LANGUAGE LEARNING

ETHEL PRUSS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

u991889

COMMITTEE

dr. Maryam Alimardani
dr. Marie Postma

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2022

STATEMENT OF CONTRIBUTION

The BCI robot-assisted learning system used for this thesis was developed in collaboration with Jos Prinsen, Caterina Ceccato, and Anita Vrins, who all contributed equally to the development of the system and were involved in the coding of all components. In terms of work division, Anita and I primarily focused on creating the experiments including code for randomly generated ROILA word lists for each condition and participant, matching randomized vocabulary tests with a graphical user interface (GUI) and interactions with the robot (including introductions and teaching the aforementioned randomized vocabulary in an adaptive manner based on user engagement and a synchronized GUI which simultaneously displayed the spelling of the words) and similarly, the adaptive screen condition involving a GUI with a virtual robot in video form and subtitles. Meanwhile, Jos and Caterina focused on creating the BCI component that formed the basis for the adaptive behavior of the robot and creating the adaptive gestures that were used by the robot. The 35 experiments conducted (8 pilots and 27 final experiments) were divided equally between the four of us with each experiment having at least two researchers present and lasting approximately 1.5 hours.

ACKNOWLEDGMENTS

I would like to thank my supervisor dr. Maryam Alimardani for all her guidance and support and Jos Prinsen, Caterina Ceccato, and Anita Vrins for their collaboration.

# THE EFFECT OF EMBODIMENT ON ADAPTIVE ROBOT ASSISTED LANGUAGE LEARNING

ETHEL PRUSS

**Abstract**

Robots are becoming increasingly popular as a teaching aid in language learning. For language learning, which relies on inter-personal interactions and references to the physical world, embodiment and the ability to adapt to the student are both important factors. In this study, adaptive behavior and embodiment were combined in a language learning experiment using a social robot as a tutor. An online passive brain-computer interface (BCI) system based on the the EEG Engagement Index (associated with attention and task engagement) was used to detect lapses in attention and prompt adaptive responses from the robot tutor, which involved additional repetition of content and iconic gestures to represent content. To isolate the effect of engagement in such a system, participants completed learning tasks in two conditions: one where the robot was physically present, and another where the robot appeared on a screen in video form, similarly to what an online Zoom class might look like. Despite no changes in behavior or content, increased learning outcomes were observed in the embodied condition, confirming that a teacher's physical presence is an important factor in learning. In addition to higher test scores, participants also reported higher engagement and more positive impressions of the robot in the Embodied condition, which indicates that the subjective learning experience, including impressions of the tutor, is also affected by embodiment. Additionally, positive correlation between engagement and test scores was observed in the Screen condition, however, this was not mirrored in the Embodied condition, perhaps due to the novelty effect of the robot. Notably, the EEG Engagement Index recorded during the learning task had no correlation with subjective engagement measures or test scores in either condition, which raises some interesting questions for future research.

# 1 INTRODUCTION

Adaptive learning is an automated and dynamic process which aims to personalize students' learning experiences (Kerr, 2016). Meeting the individual needs of students through personalization is widely believed to improve learning outcomes (Kerr, 2016). The advance of educational technology and e-learning have brought more options for adaptive learning to the table compared to classical classroom settings. Computers, robots and smart devices have been used for learning applications that automatically adjust to the user. Adaptations can be targeted at modifying content, learning paths, feedback or presentation (Martin, Chen, Moore, & Westine, 2020).

Most adaptive learning applications use subjective student feedback or the student's prior interactions with the system as a way to monitor engagement and progress levels, which in turn can be used to determine content and interventions from the system (Martin et al., 2020; Wang et al., 2020). A shortcoming of these systems is the reliance on large amounts of data for making accurate predictions about the user and the cold start problem, where in the beginning of an interaction with a new user, there is no data available (Wang et al., 2020). In recent years, an alternative approach has emerged in the form of brain computer interfaces (BCIs), which detect neural correlates of attention and engagement in order to monitor users and predict and improve learning outcomes (Yuksel et al., 2016).

EEG is an affordable and non-invasive option for recording brain activity and new wireless and dry-use devices make its use possible in a wide range of settings (Jamil, Belkacem, Ouhbi, & Lakas, 2021). The alpha, beta and theta frequency bands in the EEG signal have been connected to attention and task engagement (Coelli et al., 2015; Khedher, Jraidi, & Frasson, 2019; McMahan, Parberry, & Parsons, 2015). For example, in a 2019 study of student engagement in a VR learning environment, Khedher et al. found that the level of engagement measured by the EEG Engagement Index showed a significant positive relationship with student learning outcomes.

In theory, this should allow us to create adaptive learning systems that adjust their behavior based on the user's engagement level detected from their brain activity. For instance, the system could call for attention when it detects a drop in attention to prevent students from missing information due to lapses in attention. However, research that tests EEG-based adaptive learning applications is still very limited, particularly when it comes to robot assisted learning systems. Intelligent behavior is especially important when learning systems use robot tutors. The social robots that are often used in language learning studies mimic human behavior and appearance, creating an expectation of human-like responses. As such, when the robot

tutor does not adapt to the student's struggles with the learning material as a human teacher would, it can reduce the perceived quality of the interaction by failing to meet the student's expectations.

So far, research that explores the possibility of using BCI as a learning enhancement has been primarily conducted in virtual learning environments (Khedher et al., 2019; Rohani & Puthusserypady, 2015). This raises the question of whether the positive effects that have been found for engagement and learning outcomes would transfer from a virtual tutor to a physically present robot tutor (Khedher et al., 2019). The human-robot interaction (HRI) studies that have looked into this have found mixed results. In some instances, the physical presence of a robot tutor only seems to have an effect on subjective preference, but not learning outcomes or motivation (Kennedy, Baxter, & Belpaeme, 2015a; Looije, van der Zalm, Neerincx, & Beun, 2012). On the other hand, Köse et al. (2015) and Kennedy, Baxter, and Belpaeme (2015b) observed significant improvements in learning outcomes, motivation and engagement as an effect of physical presence, and in 2012, Szafir and Mutlu reported very promising improvements in information recall in the context of a story-telling task that combined an embodied robot tutor with BCI-based adaptive teaching. In the latter case, the results have not been replicated for language learning tasks.

The proposed study will add to this body of research by developing a novel adaptive learning system with an online EEG-based BCI component to automatically modulate learning based on engagement. The application will then be evaluated with a language learning task to investigate whether such a system can improve student learning outcomes. Furthermore, the study aims to investigate whether engagement modulating interventions are more effective when given by an embodied agent. To this end, a social robot will be used as an embodied adaptive tutor in one condition, whereas the other condition will use an adaptive desktop application with a virtual robot tutor. This leads to the following research questions and hypotheses:

RQ1 *Does embodiment have an effect on **learning outcomes** in second language learning with an adaptive robot tutor?*

RQ2 *Does embodiment have an effect on **engagement** levels in second language learning with an adaptive robot tutor?*

RQ3 *Does embodiment have an effect on the **impressions** of the adaptive robot tutor?*

## 2 RELATED WORK

Learning is a social process that involves interaction between teachers, students and peers. In recent years, technology has become more integrated

with learning but as a result, the social aspects of learning are sometimes neglected when physically and socially interactive environments are replaced with screen applications. Social robots can attempt to fill this gap, particularly when it comes to language learning, which benefits both from a social context and the ability to perform physical gestures. Several studies have shown the efficiency of adaptive robot tutors in a language learning setting (Alimardani, van den Braak, Jouen, Matsunaka, & Hiraki, 2021; Szafir & Mutlu, 2012; Wit et al., 2018). Robot tutors can fill a multitude of roles, including motivating and supporting learners (Ahmad, Mubin, Shahid, & Orlando, 2019; Donnermann, Schaper, & Lugrin, 2021; Jones, Bull, & Castellano, 2018; Liles, 2018; Ramachandran & Scassellati, 2015), and adding more dimensions to the content being taught through gestures and physical presence (Stower & Kappas, 2021; Wit et al., 2018). For instance, Wit et al. (2018) found that iconic gestures performed by a robot during a vocabulary learning task increased both long term memorization and student engagement. These studies demonstrate the promise of social robots as learning aids.

While most human learning is organized in group settings - such as classrooms and lecture halls - how much we absorb of the information presented to us and at what pace remains highly individual. A 2021 review of adaptive learning tools that personalize the learner's experience found that all but one study in the 66 reviewed reported some form of benefit to adaptive learning over non-adaptive learning (Alqahtani, Kaliappen, & Alqahtani, 2021). Despite these benefits, it would not be imaginable for one teacher to match the learning pace of each student and the infrastructure needed for adaptive learning can be an obstacle (Alqahtani et al., 2021). However, with the modern advancements of virtual online learning systems and social robots that can teach, adaptive learning has become possible at a wider scale than ever before and some studies already show promising results for adaptive learning enabled by technology (Sampayo-Vargas, Cope, He, & Byrne, 2013; Yuksel et al., 2016). For example, Sampayo-Vargas et al. (2013) used a gamified learning application to teach participants Spanish vocabulary. The game had two conditions: one where the difficulty was increased incrementally and another where the difficulty was adaptively changed based on the student's interactions with the system. The learning outcomes in the adaptive condition were significantly better. This indicates that adaptive learning systems could be a way to optimize learning.

In order to adapt to the student, we first need to understand the student's experience. Previous studies have evaluated user engagement during interactions with a robot tutor using video analysis (Wit et al., 2018) and post-interaction surveys (Donnermann et al., 2021), which can be useful

for improving the behavior of the robot in the long term. However, these methods don't allow for real time adjustments based on the individual user's mental state. A solution for this could be passive BCI, which would perform online signal analysis to facilitate adaptive teaching. This has not been explored thoroughly in the context of robot tutors, likely because measuring the level of difficulty or mastery of content through signal analysis is not an easy task. There is no single neural correlate for these concepts, rather, it is likely that they consist of a multitude of factors. Some of these factors, such as attention and engagement, can be reliably measured by EEG data and studies indicate that this could be sufficient to create adaptive robot behavior in a learning setting (Alimardani et al., 2021). In an educational setting, EEG is the most accessible BCI solution. Modern state of the art EEG headsets are wireless, comfortable and completely portable, which makes them viable for personal use, as well as being increasingly affordable (Jamil et al., 2021) and preliminary studies have successfully used EEG as a tool to monitor student engagement (Berka et al., 2007; Khedher et al., 2019).

In the context of HRI, embodiment can be seen as the difference between a physically present robot and a simulated virtual robot (physical presence and embodiment will be used interchangeably in this thesis). Most studies that investigate passive BCI for evaluating learning, embodiment has not been considered - either virtual environments (Khedher et al., 2019) or visual/auditory tasks with no tutor figure have been used (Soltanlou et al., 2019; Watanabe et al., 2016). However, several studies show that embodiment could have an impact on either student experience (Kennedy et al., 2015a; Looije et al., 2012), learning outcomes, or both (Kennedy et al., 2015b; Köse et al., 2015) and as such, it is an important factor to consider in an adaptive learning system, particularly since there is no conclusive answer on whether the physical presence of a tutor impacts learning outcomes. In the only existing example of an adaptive BCI learning system combined with an emobdied tutor, Szafir and Mutlu (2012) monitored participants' engagement during a storytelling task using an online measurement of the EEG Engagement Index, which was assumed to indicate the participants' level of engagement with the story. When a drop in engagement was detected, the robot would try to re-engage the participant by using arm gestures and increased volume. With the help of these attention-catching cues, participants' recall of the story increased 43% over the baseline (Szafir & Mutlu, 2012). The results of this study indicate that an EEG-based adaptive learning systems with an embodied robot tutor could be successful.

However, despite a number of years passing, the findings of Szafir and Mutlu have not been replicated and many possible variations have not

been explored yet (Szafir & Mutlu, 2012). For instance, it is not clear what proportion of the learning gains in the study could be attributed to the embodied agent (Szafir & Mutlu, 2012), or whether similar gains would be observed when a language learning task is used instead of a story recall task. As such, it would be interesting to see whether the results can be replicated for a second language vocabulary learning task in an experiment design that isolates the embodiment factor. Additionally, it remains to be seen if other adaptive behaviors besides attention signaling are effective in improving learning outcomes (Szafir & Mutlu, 2012), such as adjusting the learning content or repeating content as needed. Consequently, the focus on this study will be on an adaptive learning system that uses the EEG Engagement Index to assess the user's task engagement in two conditions: an embodied condition with a physically present social robot and a screen condition with a virtual robot agent. This will be tested with a language task involving learning new vocabulary in a second language. The system will adapt to the learner's needs by by repeating the content that it believes was missed due to a lapse in attention with an added iconic gesture for additional emphasis.

The EEG Engagement Index, which was also mentioned above in the context of the Szafir and Mutlu (2012) study, is a promising neural index based on EEG data in the alpha, beta and theta power bands. It has been found to correlate with states of attention and vigilance (Pope, Bogart, & Bartolome, 1995) and has been used in several studies as a way to approximate task engagement (Alimardani et al., 2021; Khedher et al., 2019; Szafir & Mutlu, 2012). Although it is not entirely clear which cognitive process this index most accurately reflects, based on the previous literature in this field, this study will use it as a measure of task engagement in the context of a language learning task.

## 3 METHODS

### 3.1 *Experimental Environment*

The adaptive learning system used for the experiment consisted of three main parts: the BCI system, the adaptive robot tutor and the screen-based adaptive learning system with a virtual tutor. The schematic overview of the system can be seen in Figure 1. Condition 1 used a custom adaptive desktop application and condition 2 will used a social robot with a custom adaptive behavior program.

A BCI system using the Unicorn Hybrid Black EEG headset developed by g.tec Medical Engineering GmbH Austria (2019) was used to collect

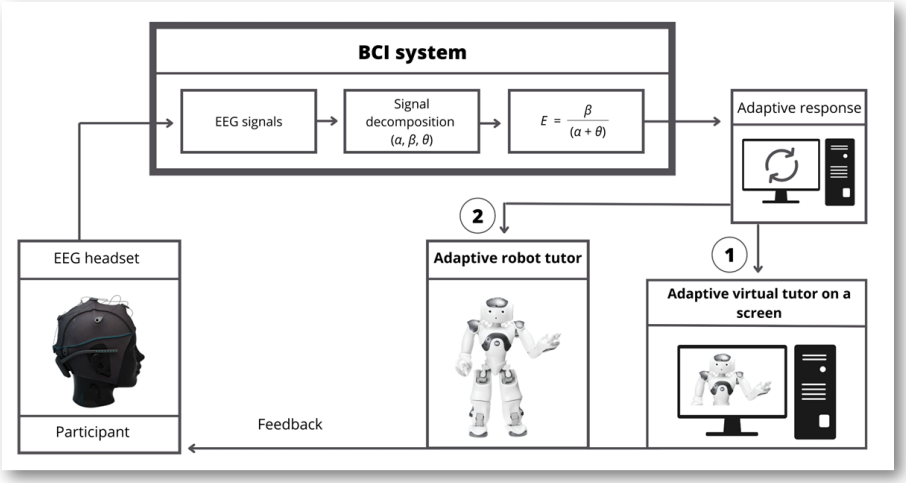


Figure 1: The closed online BCI loop for experiment 1 using the screen-based adaptive learning application, and for experiment 2 using the adaptive robot tutor.

EEG signals from the three frontal channels (Fz, F3, F4) with the electrode placement seen in Figure 2.

Pre-processing and signal decomposition was then performed to extract the relevant EEG frequency bands. First, the EEG data went through a 50 Hz notch filter and a 0.5-30 Hz bandpass. To further remove artifacts and eye blinks, clipping was applied. Next, an infinite impulse response (IIR) Butterworth Filter block was used to extract the alpha (8-13 Hz) beta (13-30 Hz), and theta (4-8 Hz) power bands (Chiang, Hsiao, & Liu, 2018). The power in these bands was averaged over the three frontal channels.

At this point, the EEG Engagement Index (E) is calculated by dividing beta with a sum of alpha and theta, as seen in Equation 1 (Pope et al., 1995).

$$EEG\ Engagement\ Index = \frac{\beta}{(\alpha + \theta)} \tag{1}$$

The EEG Engagement Index can be quite unstable and it is easily influenced by noise, so before using the calculated E value to trigger a response by the adaptive learning system, it is important to get a smoothed value. This was achieved by taking an exponentially weighted moving average (EWMA) and averaging E over 4 seconds. Additionally, E values will differ for each participant. To account for this, we created a normalized E value (Equation 2) for each participant using their minimum and maximum E values ($E_{min}$ and $E_{max}$). The $E_{min}$ was obtained during a brief calibration moment in the beginning of the experiment, where the participant was

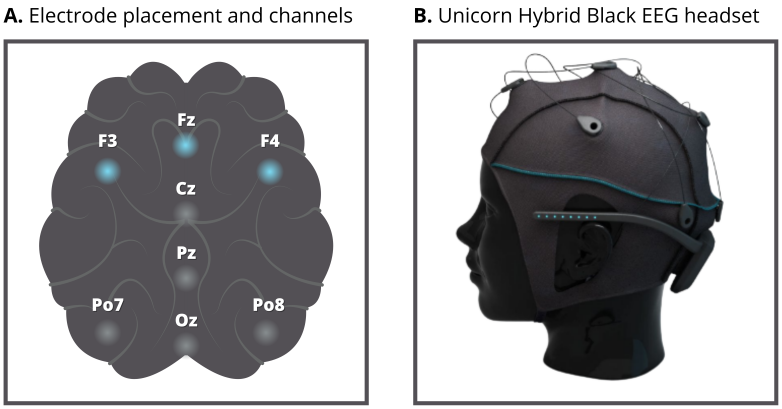


Figure 2: A. Electrode placement and B. the Unicorn Hybrid Black EEG headset used in the experiment (g.tec Medical Engineering GmbH Austria, 2019). The electrodes highlighted in blue (Fz, F3, F4) represent the channels used for calculating the EEG Engagement Index.

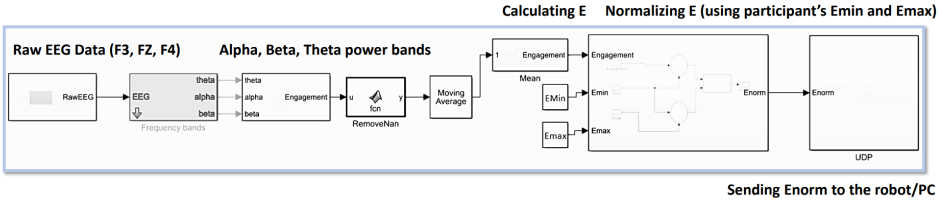


Figure 3: The Simulink model used for signal processing

instructed to look at the robot and rest. The $E_{max}$ was extracted with a calibration task performed by NAO, which was a short n-back style memory game.

$$E_{norm} = \frac{(E - E_{min})}{(E_{max} - E_{min})} \tag{2}$$

Finally, the resulting normalized E value was used to check whether an adaptive intervention is needed. If the value fell below a specified threshold (0.55), indicating low engagement or a lapse in attention from the participant, a signal was sent via the User Datagram Protocol (UDP) to the PC or robot, which then triggered an adaptive additional repetition of the last word and an iconic gesture to accompany the word. The entire signal processing process was done using Simulink, including the Signal Processing Toolbox, and the Unicorn Hybrid Black Simulink API developed by g.tec (g.tec Medical Engineering GmbH Austria, 2019). The simulink model we developed can be seen in Figure 3.

## 3.2 *Experimental Procedure*

30 participants were recruited for this experiment from the SONA Human Subject Pool system. Participants were required to be proficient in English, have normal or corrected vision and no history of migraines or seizures. All of the participants who proceeded with the experiment signed an informed consent form confirming this and agreeing to the conditions of the experiment as well as the data storage protocol (after being sufficiently informed beforehand). One participant opted not to sign the consent form and was subsequently not included in the study. Another two participants were unable to come to the experiment, leaving 27 participants ($N = 27$). All participants were Tilburg University students and were awarded 1.5 Human Subject Pool credits for their participation. Participants' ages ranged from 18 to 29 ($M = 21, SD = 3$) with 11 of the participants reporting their gender as female and 16 as male. All but 3 participants reported no prior experience with ROILA (Robot Interaction Language) and 15 had no prior experience with robots, whereas 12 participants reported having some experience with robots. The amount of languages the participants spoke ranged from 1 to 4 with a median of 3.

For the experiment, a within-subject design was used, where each participant completed both conditions with a separate but comparable set of words in ROILA (Robot Interaction Language). The experiment protocol can be seen in the diagram in Figure 4. Participants were first given an introduction form and the opportunity to ask questions, after which they were asked to sign an informed consent form and fill in a background questionnaire. At this point, the robot would be brought out to interact with the participant and introduce them to the experiment as well as the concept of EEG. Meanwhile, the EEG headset would be fitted on the participant by one of the researchers. After the EEG signal in the three frontal electrodes was confirmed to be of good quality, two calibration tasks were given to record the minimum and maximum EEG Engagement Index value. With the system calibrated, the participants would then complete 3 conditions in random order, two of which are included in this study: the Embodied and Screen condition. In each condition, the participants completed a learning task where they were introduced to second language vocabulary in ROILA with a randomly chosen set of 15 words that did not overlap between conditions. Each word was repeated 3 times with an additional repetition and gesture as an adaptation only if the EEG Engagement Index dropped. Finally, a post-test vocabulary quiz and two questionnaires (assessing engagement and robot impressions) were given after each condition. The experiment lasted approximately 1.5 hours in total with built in breaks between conditions.

ROILA was chosen due to the very low chance of participants having any familiarity with the language. As such, we can assume that any knowledge demonstrated by the participants at the end of the experiment will be as a result of the interventions. ROILA is not a natural language and does not resemble any natural languages - the language was artificially constructed with two goals in mind: making the vocabulary easily distinguishable for robots and simple for humans to learn (Designed Intelligence Group, n.d.; Mubin, 2011). The words used in this experiment were an adapted version of ROILA, where the words remain in their original form, but some of the meanings in English are changed. This was done to include more words that have easily identifiable iconic gestures that can be performed by the robot as an adaptive learning aid.

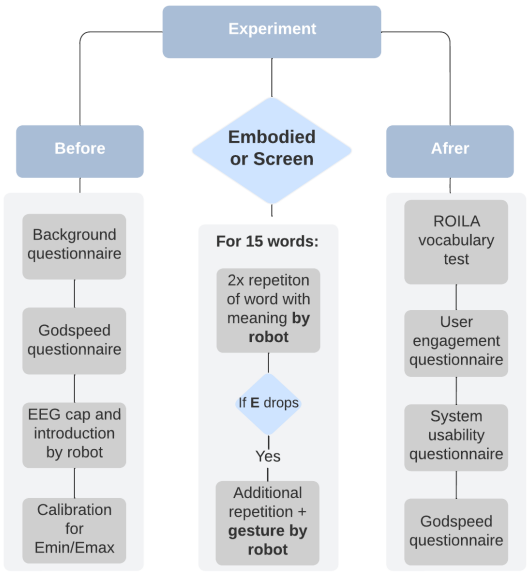


Figure 4: A diagrammatic view of the experimental protocol and the two conditions, where condition 1 uses a virtual on-screen robot tutor and condition 2 uses an embodied robot tutor. The differences between the conditions are highlighted in bold. E stands for the EEG Engagement Index, which is used to detect lapses in attention.

## 3.3 *Embodied Condition*

The NAO 6 robot running NAOqi 2.8.7 was used as the robot tutor in the Embodied condition (and recordings of the robot were used to create the Screen condition). For the Embodied condition ( Figure 5), a combination of the Python API (Python 2.7.18) and Choreographer software for NAO

(Gelin, 2017) was used to program the robot's behavior, which consisted of vocalising the ROILA words and their meanings and performing iconic gestures with an extra repetition when adaptive behavior was triggered. The words were also displayed on a screen in front of the participant simultaneously, so the participant could also see the spelling of the words. During the experiment, NAO was placed in front of the participant on a table parallel to the laptop. The robot could not be in directly the same line as the laptop due to limited space and problems with the laptop blocking gestures from the participant's field of vision or physically obstructing the robot's movement. However, we ensured that the participants could still see both the screen and the robot tutor without having to move their heads, as this would interfere with EEG data collection.

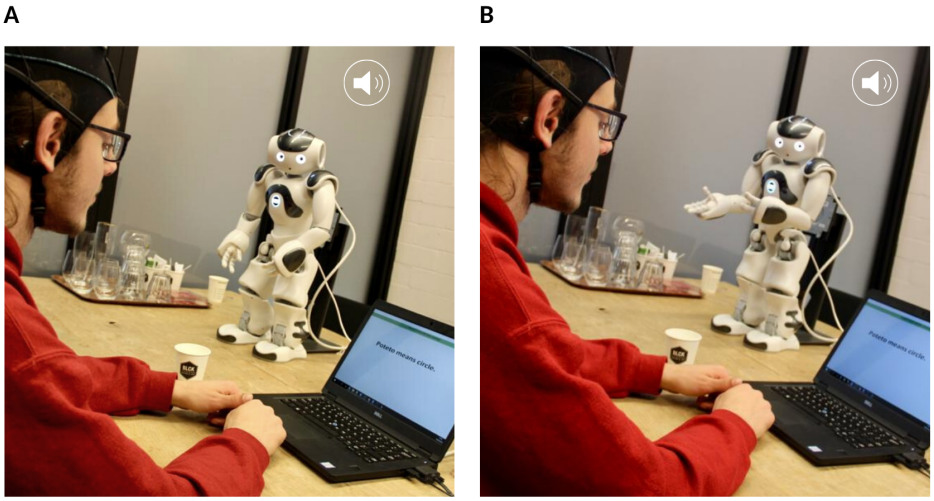


Figure 5: The Embodied condition, where A is a picture of a regular repetition (repeated twice) and B is a picture of an adaptive repetition with an iconic gesture (added as an additional third repetition if the engagement index drops)

## 3.4 *Screen Condition*

For the screen condition (Figure 6), a learning application with a virtual tutor was developed using Python 3.19. This application emulates a Zoom call, where NAO is present in video and audio form - the behavior is otherwise identical to the Embodied condition. When an adaptive response is triggered, recordings of animated iconic gestures performed by NAO are displayed in place of the live animations.

### 3.5 *Vocabulary Test*

Additionally, a vocabulary test with a graphical user interface was developed in the same coding environment. This was used for both conditions to test learning outcomes. The vocabulary test included a feature to play an audio recording of the ROILA word pronounced by NAO exactly as it was heard during the learning task, which was implemented to improve the participant's chance of being able to recognize the word in case the pronounciation and spelling are difficult to match.

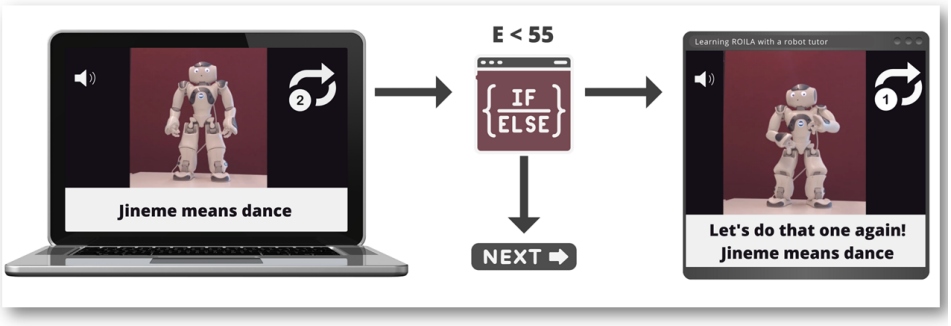


Figure 6: The Screen condition with a virtual robot tutor

### 3.6 *Questionnaires*

As mentioned above, after each condition, the participants filled three standardized questionnaires. The questionnaires are described below:

1. The Godspeed Robot Impressions questionnaire (Bartneck, Kulić, Croft, & Zoghbi, 2009), used to assess impressions of the robot tutor. The questionnaire is divided into five robot impression scales: anthropomorphism (6 items), animacy (6 items), likeability (5 items), perceived intelligence (5 items), perceived safety (3 items).

2. The short-form User Engagement Scale (O'Brien, Cairns, & Hall, 2018), used to assess subjective engagement levels. The User Engagement Scale consists of 12 items divided equally between 4 sub-scales: focused attention, perceived usability (reverse coded), aesthetic appeal and reward. The sub-scales can be combined into an overall engagement score by taking a mean of all items.

3. The System Usability Scale (SUS) (Brooke, 1996) to assess how user friendly the learning system was. SUS consists of 10 items, which are combined into a single usability score that ranges from 0-100 after the appropriate data processing steps.

The Godspeed Robot Impressions questionnaire, the short-form User Engagement Scale questionnaire (O'Brien et al., 2018) to assess subjective engagement levels and the System Usability Scale questionnaire (Brooke, 1996) to assess how user friendly the learning system was. All three questionnaires used for this experiment (Godspeed Questionnaire, User Engagement Scale, System Usability Scale) are validated standardized questionnaires and yield Likert scale data, where a number of similar questions can be summed or averaged into a concept score. The questionnaires were used with all the aforementioned scales and items, no modifications were made besides replacing placeholders such as "X application" with contextually more relevant terms, in this case "language learning application". As such, it is assumed that the scale reliability reported by the authors (Bartneck et al., 2009; Brooke, 1996; O'Brien et al., 2018) will be equivalent in this study.

## 3.7 *Statistical Tests*

The data collected for each participant included:

1. EEG recordings from the three frontal electrodes (F3, Fz, F3) as well as the EEG Engagement Index, which was calculated from the EEG data in real time.

2. Scores for each vocabulary quiz, where the percentage of correct answers was used to assess the effect of the adaptive learning application on learning outcomes in the Embodied Condition and in the screen condition.

3. Likert scale data for three questionnaires: the User Engagement Scale questionnaire (O'Brien et al., 2018), the Godspeed Robot Impressions questionnaire (Bartneck et al., 2009) and the System Usability Scale questionnaire (Brooke, 1996).

In order to explore correlations between engagement and learning outcomes, Spearman's correlation coefficient was used to compare combinations of the EEG Engagement Index, subjective engagement and vocabulary test scores. Spearman's Rho was chosen due to it's suitability for non-parametric data - some of the data, particularly the EEG Engagement Index, were not normally distributed.

For the Likert scale data collected from three questionnaires, although there is some room for debate, the most prominent papers on the topic assert that parametric tests are more descriptive even for small sample sizes and non-normally distributed data and because there are no significant differences in error rates between parametric and non-parametric tests,

analyzing means and using parametric tests is recommended over the non-parametric approach (Carifio & Perla, 2008; Norman, 2010; Sullivan & Artino Jr, 2013). Norman (2010) in particular states that for Likert data, parametric tests can be used regardless of small sample size or non-normal distribution without causing errors in conclusions. Consequently, the parametric approach was chosen and one tailed paired samples t-tests were conducted for each scale to compare the Embodied condition with the Screen condition.

# 4 RESULTS

## 4.1 *Hypotheses*

$H1_1$ *The Embodied condition will have increased learning outcomes compared to the Screen condition.*

$H1_2$ *The Embodied condition will have increased engagement levels compared to the Screen condition.*

$H1_3$ *Impressions of the robot tutor will be more positive in the Embodied condition.*

## 4.2 *Learning Outcomes*

Learning outcomes were measured by vocabulary tests conducted after each condition. The tests had a possible score range of 0-15 and 27 participants completed the tests. A Shapiro-Wilk (0.083) confirmed that the test scores are normally distributed in both conditions. Consequently, a one tailed paired samples t-test was used to compare the two conditions. The test showed a significant difference in mean scores between the Embodied and Screen condition ($t = 2.863$, $df = 26$, $p = 0.004$), with the test scores being higher in the Embodied condition ($M = 10.56$, $SE = 0.55$), compared to ($M = 8.78$, $SE = 0.60$), which supports $H1_1$.

Subsequently, the relationship between subjective engagement and test performance was analyzed using Spearman's correlation coefficient 7. Interestingly, the results show a significant correlation between self-reported engagement and test scores in the Screen condition ($R = 0.44$, $p = 0.022$). However, this does not extend to the Embodied condition, where no significant effect was found ($R = 0.094$, $p = 0.64$). On the other hand, engagement was higher overall in the embodied condition across all scores, which could be because of the novelty effect of the robot and may indicate a different type of engagement that does not benefit learning.

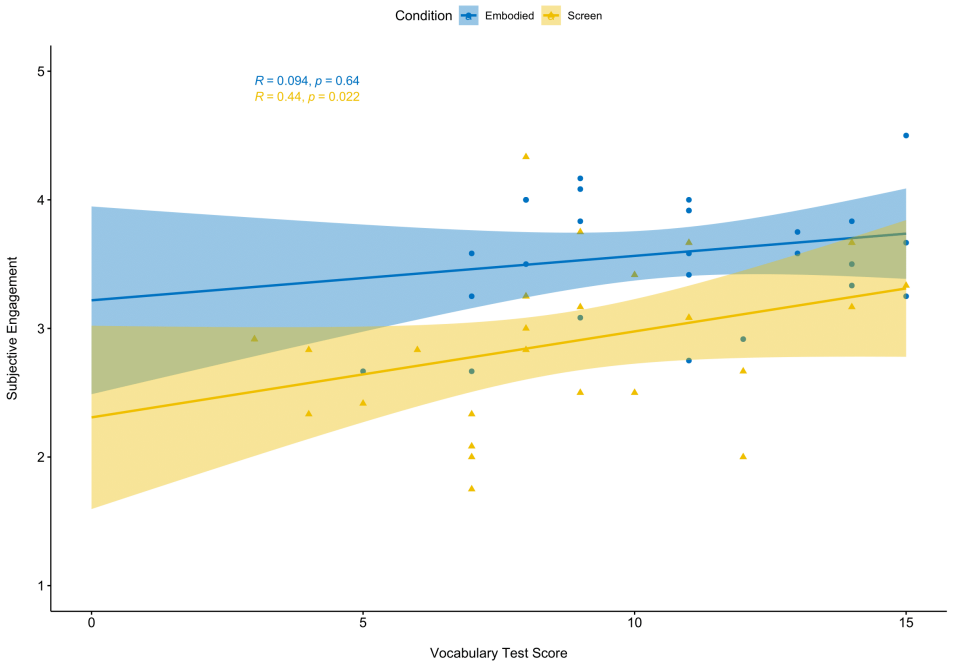


Figure 7: The subjective engagement rating from questionnaires compared to vocabulary test scores in both conditions (Embodied vs Screen).

## 4.3 *Engagement*

Two measurements of engagement were used: the EEG Engagement Index calculated from EEG data, which was recorded while participants completed the language learning tasks, and the short form User Engagemenet Scale developed by O'Brien et al. (2018), which was given after each condition. First, we will look at the EEG Engagement Index.

Due to a relatively small sample size ($N = 27$) and difficulties with calibration, the EEG Engagement Index was not normally distributed – this was confirmed by a Shapiro-Wilk test ($p < 0.001$). Additionally, before analyzing the EEG Engagement Index data, 5 outliers were removed due to excessively high normalized E values, which indicated faulty calibration; expected values after successful calibration would be roughly between 0 and 1 but due to a mix of technical difficulties with calibration and EEG data being disrupted by participants moving excessively, this was not achieved consistently. Consequently, participants with average normalized E values significantly higher than 1 ($> 1.5$) were removed. This left 22 participants ($N = 22$).

To account for the non-normal distribution, non-parametric tests were used – specifically, Spearman's correlation coefficient was used to analyze the correlations between the EEG Engagement Index, subjective engage-

ment measured by the User Engagement Scale and test scores for each condition.

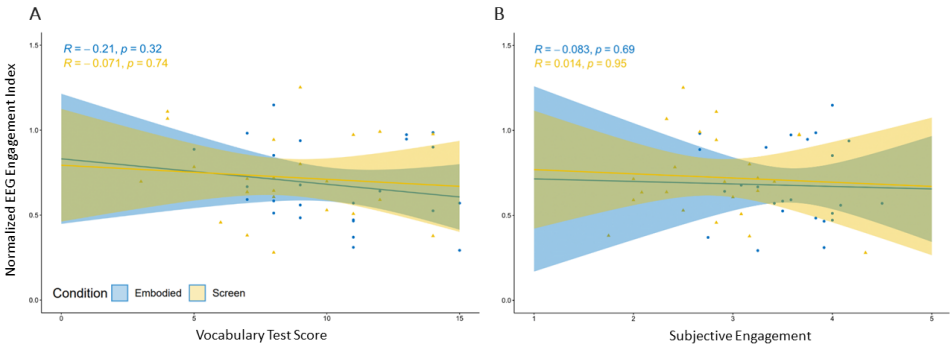


Figure 8: Normalized EEG Engagement Indices compared to vocabulary test scores (A) and subjective engagement measured by questionnaires (B) between the Embodied and Screen condition ($N = 22$). 5 outliers were removed from EEG Engagement Index data due to faulty calibration).

As seen in Figure 8 A, no significant correlation was found for either the Screen ($p = 0.74$, $R = -0.071$) or Embodied condition ($p = 0.32$, $R = -0.21$) between EEG Engagement Index and test scores. Interestingly, the Embodied condition shows a slight, although non-significant, negative trend between the engagement index and and test scores, which goes against the intuition of engagement being positively correlated with learning outcomes.

Notably, Spearman's Rho also showed no significant correlation between the normalized EEG Engagement Index and self-reported overall engagement based on the User Engagement Scale in either condition (Figure 8 B) (Screen: $p = 0.95$, $R = 0.014$; Embodied: $p = 0.69$, $R = -0.083$). This could indicate either that the EEG Engagement Index was not accurately measuring engagement or that perceived engagement and task engagement have different neural correlates or that the results were affected by issues with calibration.

Before moving on to the Engagement Scale, in this section and the ones below, Likert scale questionnaire data will be reported. As a preface before this, as discussed in the methods section, a review of best practices on analyzing Likert scale data (Carifio & Perla, 2008; Norman, 2010; Sullivan & Artino Jr, 2013) motivated the choice of a parametric statistical test for all three of the questionnaires used in this study: Engagement Scale, the Godspeed Robot Impressions Scale and System Usability Scale. All three questionnaires have been validated (Bartneck et al., 2009; Brooke, 1996; O'Brien et al., 2018) and designed for internal consistency within the scales and have the minimum informative amount of items per scale - the specifics

are discussed in more detail in the Methods section. Further principal component analysis (PCA) was not advised by the authors (O'Brien et al., 2018). The questionnaires were not modified from the originals, hence similar reliability is expected in this study. Because of the within-subject experiment design, paired samples t-tests were chosen to compare Likert scale results for every scale between the Embodied and Screen condition, where the dependent variable was the score per scale and the independent variable was the condition. The one-tailed parameter was used to test the hypothesis of the Embodied condition resulting in higher engagement and more positive impressions of the robot. Only one comparison was conducted for each scale (Embodied/Screen) and as such, no corrections were needed to adjust the p value.

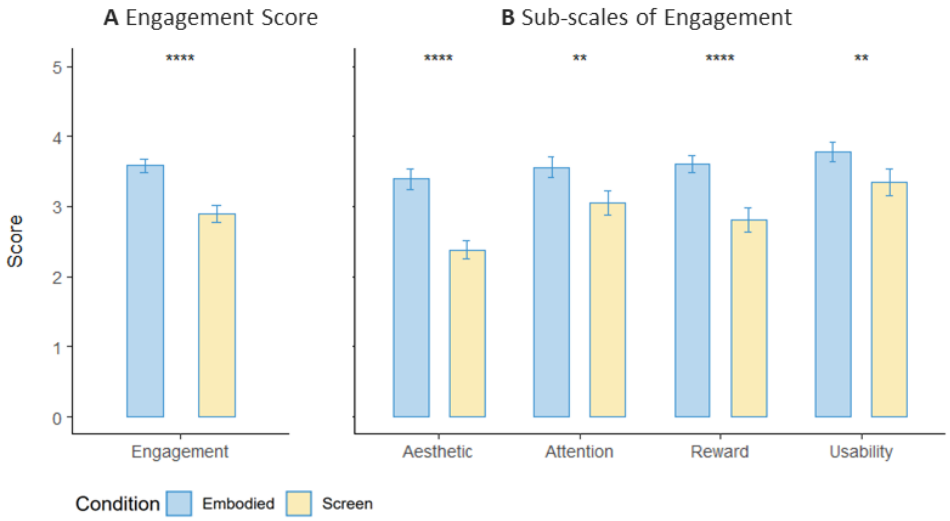


Figure 9: A comparison of the short form User Engagement Scale (O'Brien et al., 2018) results from the Embodied and Screen conditions ($N = 27$). The overall engagement score (A) is a mean of the following sub-scales: aesthetic appeal, focused attention, reward and perceived usability (B).

Figure 9 shows the results of the User Engagement Scale questionnaire in two panels. Panel A shows the over Engagement score, which is a mean of the sub-scales in panel B. Both the overall Engagement score and individual sub-scales have significantly more positive outcomes for the embodied condition (Table 1), indicating that physical presence increases student engagement during language learning tasks. The results support $H1_2$, which states that the condition with the robot tutor leads to higher student engagement compared to a virtual robot appearing on a screen.

Table 1: Comparing the User Engagement Questionnaire results between the two conditions (Embodied/Screen) using one-tailed paired samples t-tests. Each sub-scale consisted of 3 items.

| Sub-scale | Condition | $M$ $(SE)$ | $M$ Diff. $(95\%\ CI)$ | $t$ $(df = 26)$ | $p$ |
|---|---|---|---|---|---|
| Attention | Embodied | 3.56(.15) | .506(.219 − ∞) | 3.009 | .003 |
| | Screen | 3.05(.17) | | | |
| Usability | Embodied | 3.78(.13) | .432(.135 − ∞) | 2.481 | .01 |
| | Screen | 3.35(.19) | | | |
| Aesthetic | Embodied | 3.40(.15) | .012(.766 − ∞) | 7.031 | $< .001$ |
| | Screen | 2.38(.13) | | | |
| Reward | Embodied | 3.60(.13) | .802(.511 − ∞) | 4.670 | $< .001$ |
| | Screen | 2.80(.17) | | | |
| **Engagement** | Embodied | 3.58(.09) | .688(.506 − ∞) | 6.467 | $< .001$ |
| | Screen | 2.90(.12) | | | |

### 4.4 *Robot Impressions*

Figure 10 shows the results of the Godspeed Robot Impressions Questionnaire, which consists of the following scales: animacy, anthropomorphism, intelligence, likeability and safety. One tailed paired-samples t-tests were used to compare the means of each scale between the Embodied and Screen conditions. The results show that each concept scales have significantly more positive Likert scores in the Embodied condition (Table 2). The largest differences in means can be seen for the animacy, anthropomorphism and likeability impressions. Impressions of intelligence and safety are less impacted by physical presence but still show a statistically significant difference. The results support $H1_3$, which states that the condition with the robot tutor physically present in the room will lead to more favorable robot impressions compared to the virtual robot on a screen. On the other hand, no significant difference was found for the System Usability Scale rating (p = 0.278, t(26) = 0.597) between the Embodied (M = 65.83, SE = 3.34) and Screen condition (M = 64.44, SE = 3.48) and both conditions show average usability Sauro and Lewis (2016).

Godspeed Robot Impressions questionnaire, where each concept scale has significantly more positive outcomes for the embodied condition (Table 2).

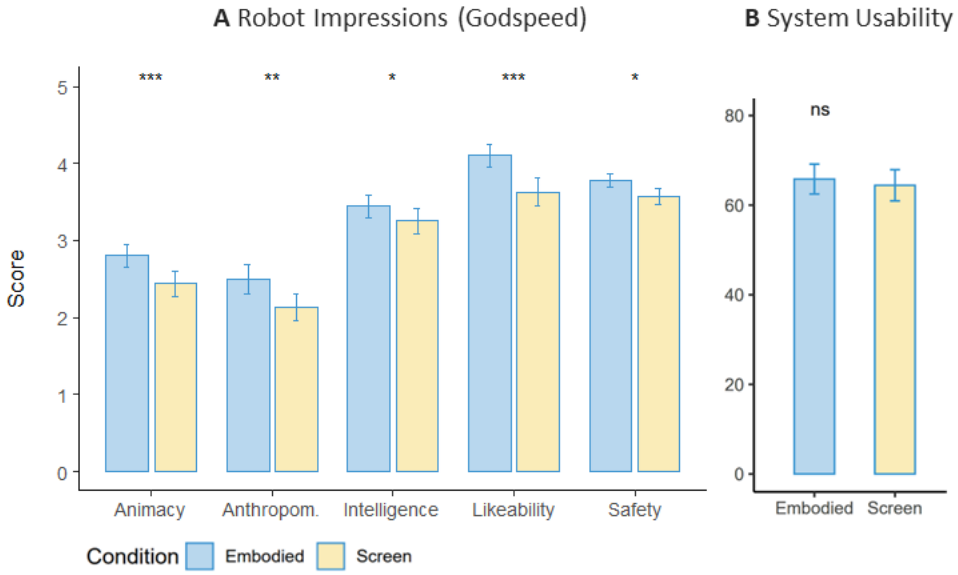


Figure 10: A comparison of the Godspeed Robot Impressions Questionnaire (Bartneck et al., 2009) results (A) and the System Usability Scale Brooke (1996) results (B) from the Embodied and Screen conditions ($N = 27$).

Table 2: Comparing the Godspeed Robot Impressions Questionnaire results between the two conditions (Embodied/Screen) using one tailed paired samples t-tests. The scales consist of 3-6 items with an average of 5; a more precise breakdown can be seen in the Methods section.

| Scale | Condition | $M$ ($SE$) | $M$ Diff. (95% $CI$) | $t$ ($df$ = 26) | $p$ |
|---|---|---|---|---|---|
| Anthropom. | Embodied<br>Screen | 2.50 (.19)<br>2.13 (.18) | .363 (.145 – ∞) | 2.843 | .004 |
| Animacy | Embodied<br>Screen | 2.80 (.14)<br>2.44 (.16) | .364 (.197 – ∞) | 3.706 | .001 |
| Likeability | Embodied<br>Screen | 4.10 (.14)<br>3.63 (.18) | .474 (.282 – ∞) | 4.212 | .001 |
| Intelligence | Embodied<br>Screen | 3.44 (.14)<br>3.25 (.16) | .193 (.018 – ∞) | 1.876 | .04 |
| Safety | Embodied<br>Screen | 3.78 (.08)<br>3.57 (.10) | .210 (.025 – ∞) | 1.935 | .03 |

# 5 DISCUSSION

The goal of the study was to assess the effect of embodiment in a robot assisted language learning setting. Robots are increasingly being used as tutors, especially for second language learning. The social robots for teaching are often humanoid in their design and mimic intelligent behavior and as such, it is assumed that students expect a degree of responsive and adaptive behavior, for example responding to their inattentiveness. The findings showed a significant positive effect of embodiment on learning outcomes (measured by vocabulary test scores). Similarly, all facets of subjective engagement were significantly higher in the embodied condition, including focused attention, how rewarding the experience was perceived as and an overall measure of engagement (using the User Engagement Scale). Questionnaire results also confirmed that impressions of the robot tutor are impacted by physical presence, which was confirmed by significantly higher ratings of anthropomorphism, animacy, intelligence, likeability and safety.

Regarding the effect of embodiment, participants experiencing higher subjective engagement and having an overall more positive impression of the learning experience and the tutor with a robot physically present is consistent with previous findings by Looije et al. (2012) and (Kennedy et al., 2015a, 2015b). The literature on the effect of embodiment on learning outcomes on the other hand was more divided. The findings from this study are in line with those reported by Köse et al. (2015), who found that embodiment had a positive effect on learning outcomes. On the other hand, it contradicts Looije et al. (2012) and Kennedy et al. (2015a) who did not find a significant effect from embodiment. It seems that benefits to learning outcomes are highly dependent on the exact experimental setup and learning task and may also be dependent on the adaptiveness of the tutor. In this, it seems that second language vocabulary tasks involving an adaptive robot tutor benefit from physical presence. Perhaps language tasks that involve iconic gestures to represent meaning (as the one in this study did) are particularly well suited for an embodied tutor, but further research is needed to untangle the exact factors that determine the presence or absence of the embodiment effect.

Additionally, there was some indication that a higher level of engagement can improve learning outcomes: subjective engagement showed a positive correlation with test scores. However, this effect was only seen in the Screen condition. The overall subjective engagement level as well as the test scores were higher throughout the Embodied condition, whereas the engagement level remained approximately the same across test scores. It is possible that the novelty effect from interacting with an embodied

robot was so stimulating that a ceiling effect was reached for subjective engagement, masking any possible correlation with the test scores. It is also possible that not all forms of engagement are beneficial for learning - for instance too much stimulation from interacting with a robot for the first time might be distracting. The latter would be similar to the findings of Kennedy et al. (2015b), who found that additional social or adaptive behavior from a robot could even negatively impact learning outcomes.

Interestingly the EEG Engagement Index did not show the expected correlation with vocabulary test scores or subjective engagement, conflicting the findings of Szafir and Mutlu (2012). No significant difference was found in the average normalized engagement values between conditions and there was also no significant correlation with test scores or self-reported engagement. As subjective engagement did show a correlation with test scores as well as differences between the two conditions, the lack of effect might indicate a few different things: the EEG Engagement Index might not be a good measure of engagement for language learning tasks (although it had a positive effect on story-based learning in the Szafir and Mutlu (2012) study), subjective experience of engagement and the more task focused engagement purportedly measured by the EEG Engagement Index might be different in terms of neural correlates and different types of engagement could have different implications for learning outcomes. Importantly, the results could also be effected by flaws in the experiment itself, particularly difficulties with calibrating the system to each participant.

The main shortcomings of the study are the questionable performance of the EEG Engagement Index (E) as a measure of engagement, difficulties with individual calibration and confounding effect from the novelty of the robot. Prior to this study, a pilot study was done to assess the adaptive learning system used in this experiment (Jos Prinsen, 2022). The pilot experiments generally validated the behavior of the adaptive robot tutoring system and we observed expected results for E in scenarios such as rest vs high intensity task or rest vs learning. However, the pilot study also highlighted the difficulty of calibrating the system to each individual user and the added complications from the novelty effect when working with robots.

Brain activity is highly individual and the same value of E can mean a completely different level of engagement based on the participant, so normalizing E based on the individual's peak (Emax) and lowest engagement levels (Emin) is necessary. However, finding the right minimum and maximum values turned out to be surprisingly challenging. Initially, an infinite runner game with increasing speed and a simple jump mechanic (the Dino game by Google Chrome) was used to find the maximum E value and a resting task was used for the minimum. However, since the

robot was not yet introduced at this point and the game differed from the learning task, this turned out to not be ideal. Combined with the novelty effect, this resulted in lower E values during calibration and then much higher values once the robot interaction started, which meant the adaptive behavior was not always triggered at the right times or was triggered too often. To remedy this, more robot interaction was added prior to the experiment and the calibration was changed to an *n*-back style memory game with the robot for Emax and looking at the robot in a resting state for Emin. *N*-back tasks have widely been used to measure working memory and attention control Kane, Conway, Miura, and Colflesh (2007), because vocabulary learning tasks likely use similar cognitive functions it was assumed that brain activity during this calibration task would more closely mimic the brain activity during the learning tasks. While this did improve the accuracy of calibration to some extent, calibration still did not work for everyone: some participants displayed higher engagement related brain activity during rest and seemed to calm down during the learning phases, which could be due to nervousness or mind-wandering and some participants moved excessively during EEG recording despite instructions to stay still, which lead to noisy data with artificially high E values.

Another concern is distinguishing between desirable levels of high engagement, such as a focused flow state, and high engagement caused by frustration or information overload. For some participants, EEG engagement levels were higher during the screen condition, but semi-structured interviews after the experiment revealed that they felt frustrated and had trouble focusing and memorizing words while learning from the screen application, which could indicate that high levels of engagement are not always positive or beneficial for learning.

What we also learned from the semi-structured interviews is that different learning strategies can impact the usefulness of adaptive interventions: participants who reported that they had a good memorization technique - for example mnemonics - sometimes found the additional repetitions and interventions from the robot frustrating as it intervened with their memorization flow, although in the few cases that this was reported the participants still scored very highly on the tests regardless of the frustration. On the other hand, participants who had no specific strategy seemed to find the adaptive interventions more helpful. For future systems, an initial calibration for individual preference might make the system more suitable for all users.

An important avenue for future work is elucidating the connection between subjective engagement, neural engagement and learning performance. While this study found some correlations between learning performance and subjective engagement, the same was not true for neural

engagement (measured by the EEG Engagement Index). In combination with a lack of correlation between subjective engagement, this raises questions about whether the EEG Engagement Index is accurate as a measure of engagement, or whether it measures a different cognitive construct than what is seen as subjective engagement. In either case, despite highly positive findings in a previous study using the same index in a story-based learning task with a NAO robot (Szafir & Mutlu, 2012), in this experiment the measure was not correlated with language learning performance, which might indicate that a different neural measure should be used for similar tasks.

Despite the difficulties, a novel adaptive BCI robot-assisted learning system was developed and validated, which shows the possibility of integrating BCI into the world of human robot interactions and robot assisted teaching as a new and improved way to personalize interactions between robot tutors and students. Additionally, the findings of this study show that physical presence plays an important role in learning, particularly language learning. Embodiment or lack thereof in education has seen drastic changes in recent years, from a complete transfer to online education with teaching moving from lecture halls to Zoom, to a gradual return to normality with some caveats. Although some students and teachers may have preferred the convenience of staying at home, it does seem that embodiment benefits the overall learning experience as well as improving learning outcomes. The results do however raise some interesting questions about the appropriateness of the EEG Engagement Index as a measure of engagement for language learning tasks and the practicalities of calibrating adaptive BCI learning systems for individual users.

## 6 CONCLUSION

Robot-assisted language learning is a popular new avenue of research. Previous research had indicated that embodiment could be an important factor in language learning and similarly, that teaching in a personalized adaptive manner can improve learning outcomes. The combination of adaptive behavior and embodiment in robot assisted language learning had not been studied much previously, although what little research there was very promising in terms of learning gains. In this study, the embodied and adaptive factors were combined in a language learning study. A novel adaptive robot assisted learning system was developed using passive BCI to module the behavior of the robot tutor such that it would intervene when low engagement is detected from live EEG data.

To isolate the effect of embodiment in such a setting, the experiment was conducted both with a virtual robot tutor and an embodied robot

tutor. Through a combination of EEG data and questionnaires, the study confirmed that embodiment has a positive effect on engagement, impressions of the robot tutor and most importantly, learning outcomes. These findings further support the important role of embodiment in learning, which is especially interesting in a time where screen-based online learning is gaining in popularity, whether it be due to necessity (due to the COVID-19 pandemic) or convenience in reduced travel times. Even with identical content and delivery, physical presence alone can lead to higher test scores and a more engaged learning experience. On another note, the findings raise questions about the EEG Engagement Index as a measure of engagement during learning tasks - further research is needed to assess what type of tasks if any the index is appropriate for and to unravel the connection between perceived engagement and its neural correlates.

## REFERENCES

Ahmad, M. I., Mubin, O., Shahid, S., & Orlando, J. (2019, 5). Robot's adaptive emotional feedback sustains children's social engagement and promotes their vocabulary learning: a long-term child–robot interaction study:. *https://doi-org.tilburguniversity.idm.oclc.org/10.1177/1059712319844182*, *27*, 243-266. doi: 10.1177/1059712319844182

Alimardani, M., van den Braak, S., Jouen, A. L., Matsunaka, R., & Hiraki, K. (2021, 11). Assessment of engagement and learning during child-robot interaction using eeg signals. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13086 LNAI*, 671-682. doi: 10.1007/978-3-030-90525-5_59

Alqahtani, R., Kaliappen, N., & Alqahtani, M. (2021, 1). A review of the quality of adaptive learning tools over non-adaptive learning tools. *International Journal for Quality Research*, *15*, 45-72. doi: 10.24874/IJQR15.01-03

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.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., ... Craven, P. L. (2007, 5). Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation Space and Environmental Medicine*, *78*.

Brooke, J. (1996). SUS: a quick and dirty usability scale. *Usability evaluation in industry*, *189*(3).

Carifio, J., & Perla, R. (2008). *Resolving the 50-year debate around using and misusing likert scales* (Vol. 42) (No. 12). WILEY-BLACKWELL COMMERCE PLACE, 350 MAIN ST, MALDEN 02148, MA USA.

Chiang, H. S., Hsiao, K. L., & Liu, L. C. (2018, 12). Eeg-based detection model for evaluating and improving learning attention. *Journal of Medical and Biological Engineering*, *38*, 847-856. doi: 10.1007/S40846-017-0344-Z/TABLES/5

Coelli, S., Sclocco, R., Barbieri, R., Reni, G., Zucca, C., & Bianchi, A. M. (2015, 11). Eeg-based index for engagement level monitoring during sustained attention. In (Vol. 2015-November, p. 1512-1515). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/EMBC.2015.7318658

Designed Intelligence Group, D. o. I. D., Eindhoven University of Technology. (n.d.). Roila vocabulary.

(https://roila.org/language-guide/vocabulary/, [Accessed 18-March-2022])

Donnermann, M., Schaper, P., & Lugrin, B. (2021). Towards adaptive robotic tutors in universities: A field study. In (Vol. 12684 LNCS). doi: 10.1007/978-3-030-79460-6_3

Gelin, R. (2017). Nao. *Humanoid Robotics: A Reference*, 1-22. doi: 10.1007/978-94-007-7194-9_14-1

g.tec Medical Engineering GmbH Austria. (2019). *User manual for unicorn brain interface hybrid black* (1.18.00 ed.). https://www.unicorn-bi.com/.

Jamil, N., Belkacem, A. N., Ouhbi, S., & Lakas, A. (2021, 7). Noninvasive electroencephalography equipment for assistive, adaptive, and rehabilitative brain–computer interfaces: A systematic literature review. *Sensors 2021, Vol. 21, Page 4754*, *21*, 4754. doi: 10.3390/S21144754

Jones, A., Bull, S., & Castellano, G. (2018, 9). "i know that now, i'm going to learn this next" promoting self-regulated learning with a robotic tutor. *International Journal of Social Robotics*, *10*, 439-454. doi: 10.1007/s12369-017-0430-y

Jos Prinsen, A. V. C. C., Ethel Pruss. (2022). A passive brain-computer interface for monitoring engagement during robot-assisted language learning. *Submitted to 2022 IEEE International Conference on Systems, Man and Cybernetics*.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the n-back task: a question of construct validity. *Journal of Experimental psychology: learning, memory, and cognition*, *33*(3), 615.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015a). Comparing robot embodiments in a guided discovery learning interaction with children. *International Journal of Social Robotics*, *7*. doi: 10.1007/s12369-014-0277-4

Kennedy, J., Baxter, P., & Belpaeme, T. (2015b, 3). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. *ACM/IEEE International Conference on Human-Robot Interaction*, *2015-March*, 67-74. doi: 10.1145/2696454.2696457

Kerr, P. (2016, 1). Adaptive learning. *ELT Journal*, *70*, 88-93. doi: 10.1093/ELT/CCV055

Khedher, A. B., Jraidi, I., & Frasson, C. (2019, 1). Tracking students' mental engagement using eeg signals during an interaction with a virtual learning environment. *Journal of Intelligent Learning Systems and Applications*, *11*, 1-14. doi: 10.4236/JILSA.2019.111001

Köse, H., Uluer, P., Akalın, N., Yorgancı, R., Özkul, A., & Ince, G. (2015, 8). The effect of embodiment in sign language tutoring with assistive humanoid robots. *International Journal of Social Robotics*, *7*, 537-548.

doi: 10.1007/S12369-015-0311-1/FIGURES/7

Liles, K. R. (2018). Ms. an (meeting students' academic needs): A socially adaptive robot tutor for student engagement in math education. *ProQuest Dissertations and Theses*, 105.

Looije, R., van der Zalm, A., Neerincx, M. A., & Beun, R.-J. (2012). Help, i need some body the effect of embodiment on playful learning. In *2012 ieee ro-man: The 21st ieee international symposium on robot and human interactive communication* (p. 718-724). doi: 10.1109/ROMAN.2012.6343836

Martin, F., Chen, Y., Moore, R. L., & Westine, C. D. (2020, 8). Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Educational Technology Research and Development*, *68*, 1903-1929. doi: 10.1007/S11423-020-09793-2/TABLES/12

McMahan, T., Parberry, I., & Parsons, T. D. (2015, 1). Evaluating player task engagement and arousal using electroencephalography. *Procedia Manufacturing*, *3*, 2303-2310. doi: 10.1016/J.PROMFG.2015.07.376

Mubin, O. (2011). Roila : Robot interaction language. doi: 10.6100/IR712664

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, *15*(5), 625–632.

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, *112*, 28–39.

Pope, A. T., Bogart, E. H., & Bartolome, D. S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological psychology*, *40*(1-2), 187–195.

Ramachandran, A., & Scassellati, B. (2015). Developing adaptive social robot tutors for children. In (Vol. FS-15-01, p. 114-116). AI Access Foundation.

Rohani, D. A., & Puthusserypady, S. (2015). Bci inside a virtual reality classroom: a potential training tool for attention. *EPJ Nonlinear Biomedical Physics*, *3*. doi: 10.1140/epjnbp/s40366-015-0027-z

Sampayo-Vargas, S., Cope, C. J., He, Z., & Byrne, G. J. (2013). The effectiveness of adaptive difficulty adjustments on students' motivation and learning in an educational computer game. *Computers and Education*, *69*, 452-462. doi: 10.1016/j.compedu.2013.07.004

Sauro, J., & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann.

Soltanlou, M., Artemenko, C., Dresler, T., Fallgatter, A. J., Nuerk, H.-C., & Ehlis, A.-C. (2019, 4). Oscillatory eeg changes during arithmetic

learning in children. *Developmental Neuropsychology*, *44*, 325-338. doi: 10.1080/87565641.2019.1586906

Stower, R., & Kappas, A. (2021). Cozmonaots: Designing an autonomous learning task with social and educational robots; cozmonaots: Designing an autonomous learning task with social and educational robots.
doi: 10.1145/3459990.3465210

Sullivan, G. M., & Artino Jr, A. R. (2013). Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, *5*(4), 541–542.

Szafir, D., & Mutlu, B. (2012). Pay attention! designing adaptive agents that monitor and improve user engagement. *Conference on Human Factors in Computing Systems - Proceedings*, 11-20. doi: 10.1145/2207676.2207679

Wang, S., Christensen, C., Cui, W., Tong, R., Yarnall, L., Shear, L., & Feng, M. (2020). When adaptive learning is effective learning: comparison of an adaptive learning system to teacher-led instruction. *Interactive Learning Environments*. doi: 10.1080/10494820.2020.1808794

Watanabe, K., Tanaka, H., Takahashi, K., Niimura, Y., Watanabe, K., & Kurihara, Y. (2016). Nirs-based language learning bci system. *IEEE Sensors Journal*, *16*. doi: 10.1109/JSEN.2016.2519886

Wit, J. D., Schodde, T., Willemsen, B., Bergmann, K., De, M., Ticc, H., ... Vogt, P. (2018). The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies.
doi: 10.1145/3171221.3171277

Yuksel, B. F., Oleson, K. B., Harrison, L., Peck, E. M., Afergan, D., Chang, R., & Jacob, R. J. (2016, 5). Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state. *Conference on Human Factors in Computing Systems - Proceedings*, 5372-5384. doi: 10.1145/2858036.2858388