AN EEG-BASED ADAPTIVE LEARNING SYSTEM WITH A ROBOT TUTOR: THE EFFECTS ON LEARNING OUTCOMES

CATERINA CECCATO

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

582202

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

ACKNOWLEDGMENTS

I would like to express my gratitude to all the people who have helped me in the process of writing this thesis. First of all, I would like to thank dr. Maryam Alimardani for all the support and the great advice given throughout the whole process. Second, I wish to thank the amazing group of people that I had the chance to work on this project with, Jos Prinsen, Anita Vrins and Ethel Pruss, thank you for this great experience. Finally, I also want to thank the friends who have been giving me moral support and precious advice while writing the thesis, I am very grateful.

STATEMENT OF CONTRIBUTION

The project and experiments described in this thesis were done with the cooperation of Jos Prinsen, Anita Vrins and Ethel Pruss. We went through the literature review process together, and we decided on the topic together. Once we started working on the setup for the experiment, Jos and I focused mainly on the development of the system. We developed the Simulink pipeline together and connected it to the robot, and then everyone was part of the process of setting up the gestures on the robot. The experiment described in this thesis is the one that Jos and I focused on, while Ethel and Anita focused on an experiment comparing an adaptive robot with an adaptive screen. Everyone gave an equal contribution in the experiments themselves. The statistical analysis was conducted separately.

AN EEG-BASED ADAPTIVE LEARNING SYSTEM WITH A ROBOT TUTOR: THE EFFECTS ON LEARNING OUTCOMES

CATERINA CECCATO

Abstract

The process of learning is a widely studied topic, especially since individuals can learn in different ways, and not everyone has the same learning method . The field of Brain Computer Interaction (BCI) and EEG recordings makes it is possible to get insight into the brain activity of different people while performing a specific task. The knowledge acquired from the EEG recordings can be used to develop adaptive learning systems, which are dependent on the brain activity of the user. This topic has already been addressed in the literature, but there is still a lot to investigate. Furthermore, literature on robots used in a learning environment shows that a robot tutor can improve learning outcomes. However, research on how to combine EEG-based adaptive learning with a robot tutor has not been widely explored. The aim of this thesis is to investigate whether an adaptive learning system, based on the brain activity of the user, and mediated by a robot tutor, can improve learning outcomes. For this purpose, an experiment was conducted. Participants were taught a series of words by a robot, and in one condition the robot was adapting to their brain activity, whereas in the other one it was not. A language test and several questionnaires were then used to evaluate the learning outcomes and the learning experience. The results did not show a statistically significant difference in test results following the adaptive condition and those following the non-adaptive one, but this research gives a good basis for further investigating the topic.

1 INTRODUCTION

Learning can be a difficult task, especially since the level of concentration can vary depending on how engaged a person is in the moment. In addition, different people can have very different learning styles, and some educational means can impact their learning outcomes more than others (Bajaj & Sharma, 2018).

The field of brain-computer interaction can be helpful in monitoring the mental processes underlying learning and the related brain activity. In particular, monitoring brain activity in a learning environment can help to understand whether learning outcomes can be improved. If that is the case, the learning environment can be improved in future sessions, or it can adapt to the brain activity in real time. This could be done by including a robot tutor in the learning environment, which would adapt the teaching method to accommodate the student's needs. Thereafter, it would be possible to get insight into how personalized teaching, with the help of a robot companion, can help improve the learning process.

In the educational field, it is well known that differentiated learning can help improve learning, as it helps students being more focused with less effort (Morgan, 2014). However, for many students it is hard to find the best learning approach for their needs. Having a BCI system that automatically detects when a student is starting to struggle with a topic and is able to adapt to the needs of the student can make the learning process more fluid. The BCI component would help in making the process more personalized in order to optimize it. Furthermore, by adding a robot tutor to the system, the social component can be introduced in order to get more insight into how a personalized tutor can help improve the learning process. All the components together can be valuable contributions to the field of education and to the research on differentiated learning.

In addition, addressing this problem can be helpful in investigating how well cognitive load can be measured by monitoring brain activity when performing specific tasks. In their overview of knowledge on cognitive load theory Sweller, van Merriënboer, and Paas (2019) mention that self-regulated learning can be important but difficult to achieve. The aforementioned BCI system supported by a robot tutor can give insight on ways to implement self-regulated learning. That would be possible by focusing on the brain activity related to engagement levels when performing a specific task.

In this research, adaptive learning and a robot tutor will be combined in the attempt to answer the question:

> *Can a robot tutor that adapts to a person's brain activity improve learning?*

This will lead to valuable insight not only on adaptive learning in general, but also on how it can be applied to robots for a possible future teaching environment. In order to get insight into the topic, the following research questions will be addressed:

RQ1 *Can EEG engagement index be used as a predictor of test results in a language learning task?*

RQ2 *Can the number of interventions performed by a robot tutor influence the test score?*

RQ3 *Does an adaptive robot influence the self-reported user engagement?*

In the next section, more information on the literature behind this research will be given. That will be followed by the methods, in which the developed system will be outlined and the experimental setup will be discussed more in depth. In the results section, the outcome of the experiments will be described, and it will be further developed in the discussion. The latter section will also include reflections on the limitations of the research. Finally, the last section will sum up all the findings and draw the conclusions.

2 RELATED WORK

Learning is a widely studied topic, and it can be very complex. There are many approaches to learning and due to all the different components needed to learn, there are multiple areas that can cause issues. As shown by studies on people with ADHD and learning disorders, attention is an important part of learning. Rohani and Puthusserypady (2015) give an example of a BCI method, based on neurofeedback, that could help improving attention, and therefore obtain an enhancement in learning. The experiment was conducted on healthy participants, showing that improving attention impacts positively the learning outcomes.

However, attention is not the only component that influences learning. In fact, engagement also plays an important role. Khedher, Jraidi, and Frasson (2019) pointed out that monitoring the level of engagement in learning tasks can be very useful to understand the underlying learning process. They also found out that the level of engagement seems to be directly related to the performance in a learning task, and that this is not necessarily related to the level of attention given to a specific element.

Coelli et al. (2015) showed that EEG recordings can be an efficient tool to monitor engagement when participants are involved in a task measuring selective attention. Therefore, there is room for further investigation on whether there is a way to use information from the EEG recordings to optimize user engagement in a task. Ewing, Fairclough, and Gilleade (2016) presented an example of an adaptive game that aimed at maximizing the player's engagement. They measured brain activity focusing on indicators of effort and game demand. Therefore, four states were identified: zone, engagement, overload, and boredom. These states were then used as a basis to classify the current state of the participant during a Tetris task. This in turn made it possible to modify the difficulty of the game based on the state that the participant was in. The study provides an example of how EEG data can be used to adapt the environment to the user's engagement in real time. However, the four states identified and the whole system were strictly related to the environment of a game.

Alternatively, a similar approach in the field of learning was applied by Mohamed, Halaby, Said, Shawky, and Badawi (2019) . They proposed an experiment to detect students' cognitive states in order to create an adaptive learning environment, focused on the specific profile of each student. In this case, the adaptive learning task was only developed after the data collection, meaning that it was not a real-time adaptive learning environment. However, it is already a good example of a way in which brain activity can be used to improve the learning process. Similarly, Szafir and Mutlu (2012) proposed an experiment based on BCI in which the level of attention of the participants was monitored in real time by analysing the EEG levels for alpha, beta and theta frequencies. In this case, the learning experience was intermediated by a social robot. When a drop in attention was detected, the robot would perform a specific action to bring the attention back. This experiment shows that an adaptive learning environment helped by the presence of a robot can improve learning, and it also gives input on how different types of feedback might influence the learning process.

The previous study introduced the robot component with a story telling task, but different experiments on social robots have focused more on second language learning tasks. For example, de Wit et al. (2018) proposed a study to investigate the ability of a robot to teach a second language to children through words memorization. The results of this study showed that children manage to learn the new words and that the presence of the robot seems to improve long-term memorization. An adaptive approach was also used in this study, but it was based on knowledge, and the results were inconclusive. More studies on the use of social robots for learning tasks have outlined some of the advantages and reasons why social robots can be considered a valid option for this task. van den Berghe, Verhagen, Oudgenoeg-Paz, Van der Ven, and Leseman (2019) highlight two main advantages: the possibility to interact with a physical agent, which provides a level of grounding that is important for language learning, and the more natural interaction that derives from a human-looking agent.

However, some studies have shown contrasting results. Vogt et al. (2019) observed that children interacting with a robot did not seem to have higher learning outcomes than children interacting with a screen. Nonetheless, it is important to note that in this experiment the interaction with the robot was always mediated by a tablet that the children were interacting with, which might have influenced the results. Another experiment by Alimardani, Braak, Jouen, Matsunaka, and Hiraki (2021) showed similar results. In this study children were given a language learning task and their engagement was measured with EEG data. The performance of children interacting with a robot was comparable to that of children interacting with a display, but the level of engagement seemed to be overall higher for children interacting with the robot. This implies that a high engagement is not always a signal of higher learning performance.

The latter statement seems to contradict the body of literature indicating engagement as a valid measure to be used to improve learning outcomes. In particular, it seems to contradict the experiment by Szafir and Mutlu (2012), who also combined the measure of engagement with a social robot. However, previously mentioned, the type of task performed in that experiment was different. Combining an adaptive setting, based on EEG, and a social robot acting as a tutor in the context of second language learning could bring more clarity on whether this system can be useful for second language learning. Furthermore, it could bring valuable insight on whether engagement can be considered a good measure to track in order to improve learning outcomes in this type of task. Employing a second language learning task can have several advantages. For example, the outcomes can be confronted with the body of existing literature on second language learning, considering research where social robots were also involved, and research where the engagement levels were measured. Furthermore, this approach can also be helpful to test the learning outcomes, as most studies reviewed so far have used a language quiz administered after the interaction with the robot as a means of evaluating the learning performance (Alimardani et al., 2021; de Wit et al., 2018; Vogt et al., 2019).

As presented in this section, the literature on real-time adaptive learning based on EEG data, helped by the presence of a robot tutor, is very limited. The experiment by Szafir and Mutlu (2012) presents a good example of how this type of experiment can be performed. The aim of this thesis is to investigate whether an experiment with a similar setup, with a robot adapting to the engagement level of the participant, can lead to improved learning experience and learning outcomes.

3 METHODS

The experiment consisted of a robot tutor teaching a set of words to the participants, who were wearing an EEG cap that allowed to monitor their brain activity.

3.1 *Participants*

Data from 27 participants was collected, 16 males and 11 females. Everyone was given an information letter before the experiment started, with all the information on the procedure, possible risks, and privacy concerns. The participants were then instructed to sign a consent form if they agreed to participate to the research. Thereafter, a background questionnaire was administered, which showed that they were all between 18 and 29 years old. The mean age of the participants was 21.07, with a standard deviation of 2.99. The results of the skewness (1.1341) show that the age was slightly skewed to the right. A potentially important factor that was taken into account is the number of languages spoken by the participants. The results showed an average of 2.5 languages known, with one participant knowing 1 language, one knowing 4, and the rest falling in either 2 or 3 known languages. Data on previous experiences with robots and with ROILA were also collected. 12 participants had previous experiences with robot, but only 3 had experience with ROILA.

3.2 *BCI System*

A system was developed to extract the EEG engagement index in real time and send it to the robot. In the Simulink pipeline (Figure 1) the EEG data from the 3 frontal channels (Fz, F3, F4) was selected, and three IIR Butterworth filters were used to extract alpha (7-13 Hz), beta (13-20 Hz) and theta (4-7 Hz) band powers. The original values were squared and the mean values, averaged over channels, were taken to calculate the EEG engagement index using the formula by Pope, Bogart, and Bartolome (1995):

$$\frac{\beta}{\alpha + \theta}$$

This index was then normalized by taking the lowest engagement index, and highest engagement index, calculated during the calibration phase. After the normalization the index was sent to the python environment that was used to manage the robot through UDP. The robot was meant to teach ROILA, a spoken language for robots developed by the Department of Industrial Design at Eindhoven University of technology (Mubin, 2011). It did so by repeating a word in English and ROILA twice, and then based on the data coming from Simulink it would either repeat the word a third time, while performing an iconic gesture, or it would proceed to the next word. More information on the process can be found in the procedure section.

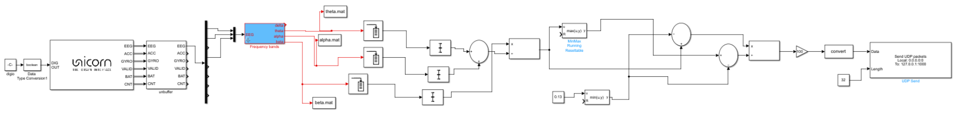

Figure 1: Simulink Pipeline

3.3 *Materials*

The Unicorn Hybrid Black EEG headset (g.tec neurotechnology GmbH, Austria) was used to extract the EEG data in real time. In particular, the Unicorn Simulink interface was used to input the data in MATLAB Simulink. The EEG headset acquires EEG data from 8 electrodes (Figure 2), and it is sampled at 250 Hz per channel with 24-bit resolution. It can be used to record data with or without gel, but in order to obtain the highest accuracy possible gel was used in the experiment. The NAO robot, developed by SoftBank Robotics, was used for the robotic component of the experiment. Nao has 25 degrees of freedom, which was useful to perform the gestures that were needed as a feedback method. Moreover, the speaking task was made possible by the use of directional microphones. Nao has a few languages already installed, but in this experiment only English was used. This means that the words in ROILA were pronounced with an English inflection, and the translations were in English.

A variant of ROILA was used in this experiment. The vocabulary consisted of a set of words that are meant to be easy to pronounce for robots, and do not resemble any human language. Thirty English words were selected and associated with 30 words in ROILA. The English words selected were chosen based on how clear the corresponding gesture made by the robot was. In cases of words with similar gestures, one of the two words was replaced. The main application used to set up the experiment is MATLAB Simulink. The pipeline described above has been created using a combination of blocks included in Simulink's DSP system Toolbox and blocks provided by g.tec, through the Unicorn Simulink Interface. Choreographe Suite was used for the gestures of the robot. Thanks to the recording function, customized gestures were created and uploaded to the robot, so that they could be executed through a python script. Python 2.7 was used for the parts where the robot was involved. Python 3 was used for other parts, such as the script to change the file names after each run in Simulink. RStudio 2022.02.2 was used to perform the statistical analysis which followed the data collection. In particular, a series of comparison and correlation analyses were performed on the data in an attempt to answer to the research questions.

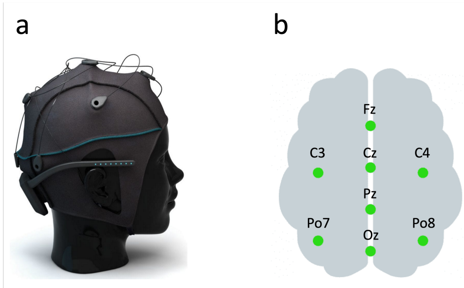

Figure 2: a) Unicorn EEG headset b) Electrodes position (Pruss et al., 2021)

3.4 *Procedure*

During the experiment, an EEG cap was applied on the participant, while the robot gave an initial presentation on what the EEG cap is and what would happen in the experiment. After that, the first step of the experiment consisted of a calibration task. This task was needed to establish the minimum and maximum engagement, used to set up the normalization step in the system. Results from the pilot studies highlighted the need for a calibration task in which the robot was already present (Prinsen, Pruss, Vrins, Ceccato, & Alimardani, 2022). Therefore, the calibration task consisted of two parts. The first one in which the participant was instructed to look at the robot in front of them while trying to relax and sit as still as possible, and the second one, in which the participant performed a short n-back task, given by the robot. The n-back task consisted of the robot asking the participant to memorize a set of 10 words in order, followed by a few questions on which word was at a certain position. The participants were instructed to only think of the answer, without saying it out loud, so that movement could not influence the EEG data. This task was chosen as it could be related to the memory task in the actual experiment. Having a calibration task that is similar to the main task was a way to ensure that the values selected would be more reliable.

After having established the values needed for the normalization, the experiment continued to the language task. Two conditions were present for each participant. They will be referred to as adaptive and non-adaptive condition from now on. For both conditions the robot taught 15 words. In the adaptive condition, the robot would say a word in English and its translation in ROILA twice, and when an engagement lower than 55 was detected, it would repeat the word a third time, while also performing an iconic gesture. In the non-adaptive condition, the third repetition was randomized, with a 15% chance of being triggered. Moreover, in the adaptive condition the third repetition was preceded by a short sentence

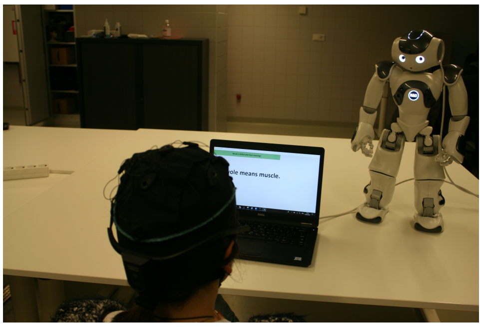

Figure 3: Experimental setup with EEG, laptop, and robot

said by the robot to acknowledge the drop in engagement. This element was removed in the non-adaptive condition. The value of the engagement index used to determine whether a third repetition was needed was an average of the indexes detected during the first two repetitions of the word. The 55 threshold selected for a third repetition to be triggered in the adaptive condition was determined after the results from the pilot studies. In both conditions, a laptop was put in front of the robot so that the participants could read the words in ROILA and English (Figure 4).This was implemented to facilitate the word recollection and because the spelling of the word was not always straightforward when only listening to the robot. The two conditions were randomized, meaning that in some experiments the non-adaptive condition was first, followed by the adaptive one, and in other ones it was the opposite. This was done to ensure that the results would not be skewed because of a novelty bias. The ROILA words that were taught to different participants were also be randomized, with the only constraint that words that appeared in one condition could not appear in the other condition. This is because of the nature of the experiment. Since it was a within-participant study, presenting the participants with words they had already seen could have biased the results. The participants were then given a quiz on a laptop to test their learning at the end of each condition. The quiz consisted of a set of a multiple-choice questions, covering all the words learnt in the previous condition. The participants could see and hear a word in ROILA and they had to select the correct translation of the word in English. There was no time limit, nor a limit in the amount of times the world could be repeated, so each participant could decide how much time to spend on the task and on each word. During the administration of the quiz, the robot was in front of the participants, behind the laptop, but it was not interacting with them

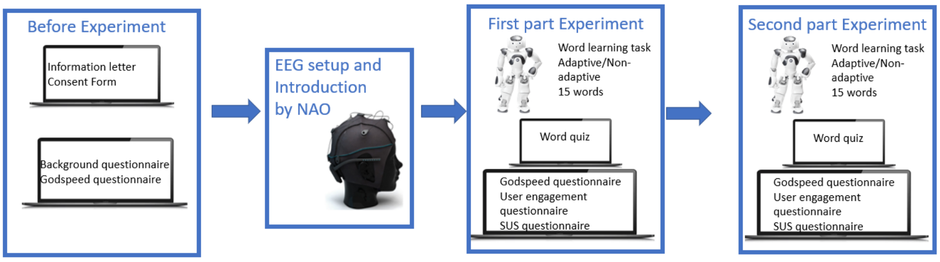

Figure 4: Diagram of the experimental procedure

3.5 *Evaluation*

The main task used to evaluate the learning outcomes was the language test described above. A pronounce button was given so that the participants could hear the robot's pronunciation of the word. In addition, several questionnaires were administered to test the user experience. One background questionnaire was already mentioned before. It was administered before the experiment, and before the participant saw the robot. This questionnaire included general demographic questions about the participants and questions on whether the participants had ever interacted with a robot or ROILA before. Moreover, a Godspeed questionnaire was used to evaluate the participant's general impression of robots (Bartneck, Kulić, Croft, & Zoghbi, 2009). The participants had to give a number on the 5-point Likert scale to judge anthropomorphism, animacy, likability, perceived intelligence, and perceived safety of the robot. The same questionnaire was also administered after each condition, applied to the participant's experience with the robot in that specific condition. In addition, the questionnaire after each condition also included the System Usability Scale (Brooke, 1996) and the short version of the User Engagement questionnaire (O'Brien, Cairns, & Hall, 2018), in which participants were instructed to fill in a number on the 5-point Likert scale. The System Usability Scale is used to measure the usability of the system, whereas the User Engagement questionnaire measures self-reported user engagement. The System Usability Scale presents several questions that are then averaged together and scaled to get a score out of a 100, which is what is then used in the analysis. The User Engagement Scale can be divided into four subscales: aesthetic appeal, focused attention, perceived usability, and reward. These subscales can also be averaged together to obtain the overall perceived engagement score.

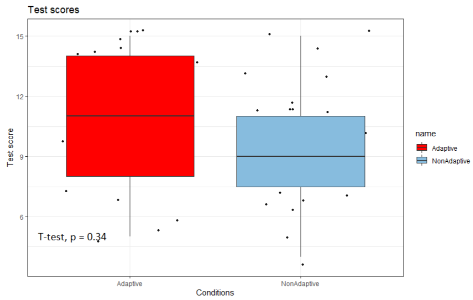

Figure 5: Comparison Test Scores - Adaptive vs NonAdaptive

4 RESULTS

4.1 *Test Results*

The word quizzes administered after each condition resulted in a score that could be between 1 and 15, with the number corresponding to the amount of correct guesses. The test results from the adaptive condition ($M = 10.56, SD = 2.87$) compared to the non-adaptive condition ($M = 9.74, SD = 3.35$) with a paired t-test did not show significant effects for condition ($t(26) = 1.2, p = 0.24$). (Figure 5). In order to address whether higher engagement levels measured by the EEG engagement index led to better test results, a correlation between normalized engagement and

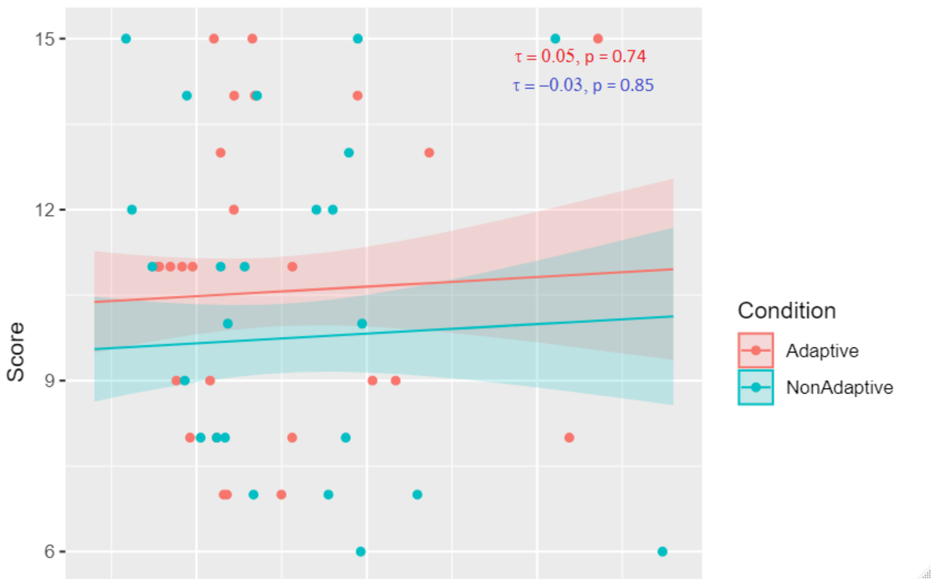

Figure 6: Correlation between engagement and score per condition

number of correct answers for each condition was done. In the adaptive condition, EEG engagement index and test scores did not show a significant correlation ($\tau = 0.05, p = 0.74$) (Figure 6). In the non-adaptive condition, no significant correlation could be found either ($\tau = -0.03, p = 0.85$) (Figure 6). Furthermore, a comparison of non-normalized EEG engagement index between conditions was performed. The non-normalized EEG engagement index for the adaptive condition ($Mdn = 0.73$) compared to the non-adaptive one ($Mdn = 0.68$) with a Wilcoxon test did not show a significant effect in this case either ($p = 0.79$).

In regards to the effect of the number of repetitions, a first analysis of number of repetitions per condition was done. Data on the amount of repetitions for 3 participants was missing, therefore the data relating to them was deleted for this step. The results of the Shapiro-Wilk normality test showed that the amount of adaptations in the adaptive condition showed a lack of a normal distribution ($p < 0.05$), whereas the amount of adaptations in the non-adaptive condition resulted in a normal distribution ($p = 0.1$). Therefore, the median was considered to compare them, and the adaptive condition showed a higher median($Mdn = 4.5$) than the non-adaptive one ($MDN = 2.0$). The third repetition in the non-adaptive condition was set to be triggered with a 15% chance, and the data showed that in this experiment it triggered 16% of the times, whereas the adaptive condition triggered 36% of the times. A test to measure the correlation

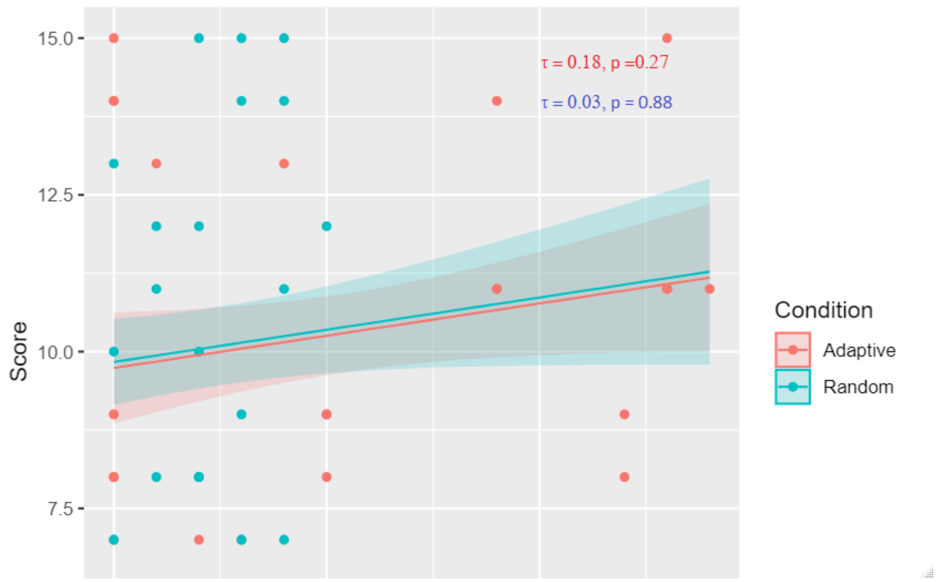

Figure 7: Correlation between test scores and amount of third repetitions - Adaptive and Non Adaptive

between amount of third repetitions and test scores for each condition was then performed. In the adaptive condition the two variables did not show correlation ($\tau = 0.18, p = 0.27$), and no significant correlation was observed in the non-adaptive condition either ($uptau = 0.03, p = 0.88$) (Figure 7).

4.2 *Questionnaires Results*

4.2.1 *Godspeed Questionnaire*

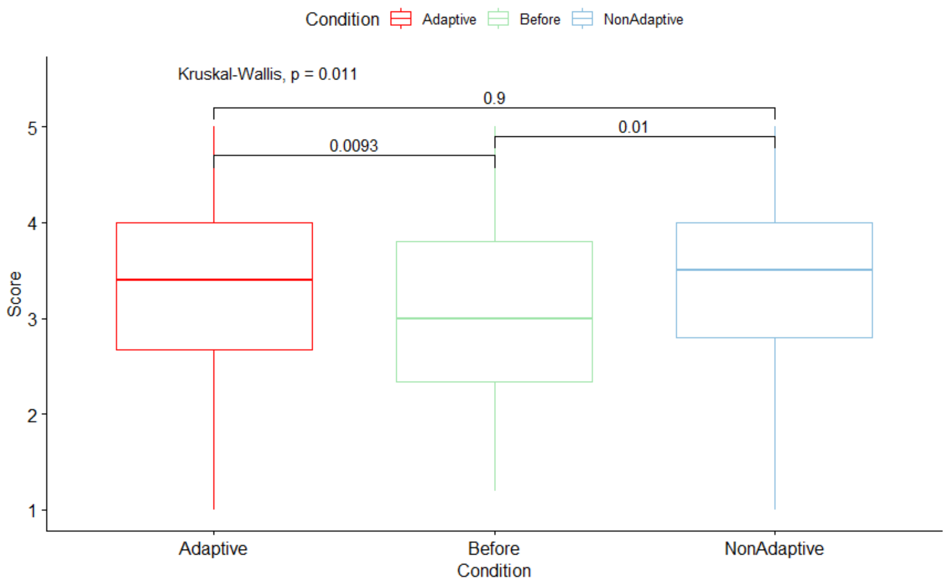

Figure 8: Comparison Godspeed general results over conditions (Non-adjusted p values)

The Godspeed questionnaire on general impressions of robots was administered before interacting with the robot and after the adaptive and non-adaptive conditions. The Shapiro-Wilk normality test was significant for all three conditions (respectively, $p < 0.05, p < 0.05 and p < 0.005$), meaning that none of the conditions had a normal distribution. A Kruskal-Wallis H test showed that there was a statistically significant difference in score between the different conditions, $(\chi)2(2) = 8.960, p < 0.05)$ (Figure 8. A Pairwise comparisons using Wilcoxon rank sum test (p value adjustment method: Bonferroni) showed that there is a significant difference in score between the before and adaptive condition ($p < 0.05$), as well as between the before and non-adaptive condition ($p < 0.05$). No significant difference in score was observed between the adaptive and non-adaptive condition. In order to investigate the comparison between conditions for the separate components, a pairwise comparison using Wilcoxon test (p value adjustment method: Bonferroni) was conducted for each concept. The test showed significant differences in score only for the likeability concept, between before and non-adaptive $(p < 0.01)$, with the non-adaptive condition $(Mdn = 4.5)$ having generally higher scores than the before condition $(MDN = 3.7)$. With Bonferroni adjustment, none of the other concepts had statistically significant results when taken separately.

4.2.2 *User Engagement Questionnaire*

O'Brien's User Engagement questionnaire consists of four components, averaged together to get the self-reported user engagement. A comparison between the different components and the average self-reported engagement for each condition (Figure 9) shows that the different components do not seem to have significant results in terms of relation between component and condition. This factor is reflected in the engagement component. Looking more in depth at the engagement component, in the adaptive condition the Shapiro-Wilk normality test showed a non-significant result for the adaptive condition $(p = 0.29)$, which confirms the assumption of linear distribution. However, in the non-adaptive condition the result was significant $(p < 0.05)$, rejecting the hypothesis of linearity, so a Wilcoxon rank sum test was performed. The results from the test showed a non-significant $(p = 0.091)$ difference in the score of the engagement component between adaptive and non-adaptive condition. However, the plot seems to indicate that the average self-reported user engagement is high for both conditions $(Mdn = 3.58)$ on the Likert scale

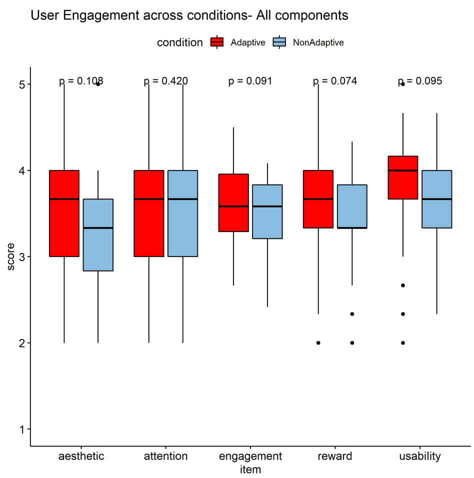

Figure 9: User Engagement results for all components

4.2.3 *System Usability Scale*

The results from the System Usability Scale questionnaire show a normal distribution of the data for the score in both conditions, confirmed by the Shapiro-Wilk normality test both for the adaptive $(p = 0.16)$ and non-adaptive $(p = 0.14)$ conditions. This led to an evaluation by means of a paired t-test, which gave non-significant results $(p = 0.26)$ on the comparison between results of the adaptive and random conditions. The boxplot generated to compare the two conditions (Figure 10) does not seem to show a relevant difference between the two conditions either. However, the general score seems to be relatively high for both conditions.

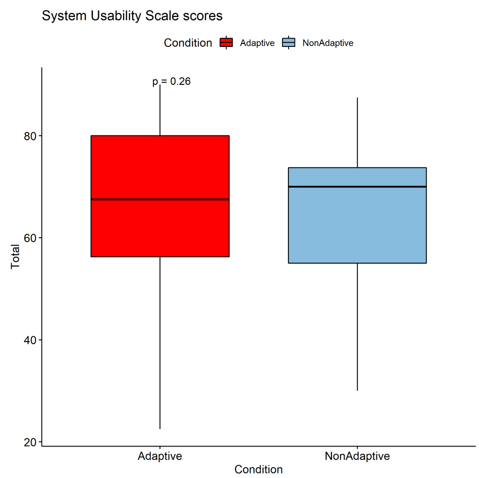

Figure 10: SUS scores across conditions

5 DISCUSSION

The experiment presented in this thesis combined a Brain Computer Interaction system monitoring the engagement level of a person, based on the EEG engagement index, with a social robot acting as a language tutor which adapts to the level of engagement of the participant. The goal was to investigate whether this adaptive approach would have an impact on the learning outcomes, leading to an improved learning process with the help of the adaptive robot. In doing so, the effect of the engagement levels on the results in a language learning task was investigated, as well as the relation between amount of interventions and test scores. Moreover, the perception of an adaptive robot compared to a non-adaptive one, as reported by the user engagement questionnaire, was considered.

The learning outcomes of the participants were measured with a language quiz administered after each condition. As the results have shown, there was no significant effect of the condition on the results in the language test. This would naturally lead to the conclusion that the initial hypothesis is wrong, but there might be several factors influencing this outcome. First of all, the learning task chosen might not be optimal to test learning outcomes. The task might have focused more on the mnemonic aspect of learning, which is certainly an important part, but there are more aspects that could be considered as well. In fact, in the case of the research by Szafir and Mutlu (2012), the adaptive robot seemed to lead to an improved performance, and one major difference between their experiment and the one presented in this report is the type of learning task performed. The participants in their experiment did not need to memorize words, but just the details of a story narrated by the robot. That type of interaction might have benefitted more from the use of a social robot, while the word recall task proposed in this research might have not used the potential of a robot tutor at its best. Furthermore, the way the task was assessed was also quite different. In the experiment by Szafir and Mutlu (2012), the robot was also part of the testing, by asking the participants questions about the story, whereas in this experiment the test was done on a laptop, with the possibility of hearing the pronunciation of the words again, but without interacting with the robot at all. The difference between the task and the testing might have influenced the test results.

Second, the EEG engagement index might not be optimal to measure learning. In fact, the results from the correlation between level of engagement and number of correct answers were also not significant, which means that the EEG engagement index cannot be used as a predictor of performance. This answers the first sub-question posed at the beginning: the results of this experiment indicate that the level of engagement as measured by the EEG engagement index does not necessarily correlate with the test scores in this language learning task. This confirms the results of the research by Alimardani et al. (2021). They showed that despite the increase in engagement measured by the EEG engagement index when children interacted with a robot, the outcomes of the learning task were comparable to those of children interacting with a screen. However, considering the fact that Szafir and Mutlu (2012) also used EEG engagement index, and they did get better results for the adaptive condition, it might also be the case that EEG engagement index is not fit to measure learning outcomes specifically in second language learning tasks. Future studies could focus on whether the EEG engagement index really is a good measure of engagement during a learning task, and on whether the EEG engagement index can be used as a predictor of learning outcomes in different learning tasks.

Third, the adaptive component of the robot consisted in a third repetition, accompanied by a gesture and a short sentence to point out that the participant's engagement had dropped. This interaction might not be the most effective way to adapt the teaching method. A more subtle method, that adapts without informing the participant about the drop in attention, might be a good alternative, possibly leading to better results. Moreover, the results can be related to the finding of Vogt et al. (2019) about iconic gestures not having a significant impact on learning outcomes. Perhaps implementing the gestures with every repetition, to keep the social aspect of gestures, and keeping only the third repetition of the word as adaptation could lead to different results.

In addition, the core reasoning behind the adaptive condition was to keep the participant in a state of flow, making sure that the engagement was constantly high, in juxtaposition with the non-adaptive task, where the average engagement would supposedly be lower. As the results have shown, this hypothesis could not be confirmed either. Even the non-normalized engagement index did not result in being higher for the adaptive condition. On the contrary, the level of engagement seemed to be comparable. This might be a consequence of a fault in the system. In fact, while doing the experiment it was observed that at times the average engagement index would drop or rise significantly from one condition to the other, even without the interventions being triggered. This is probably due to a series of factors. First, some participants reported after the experiment that when switching from one condition to the other they changed strategy to memorize the words, which might have influenced the engagement index. Second, probably the most important factor, the calibration of the system was far from optimal.

The calibration was a fundamental step to determine the normalization of the EEG engagement index. It was done to find the minimum and maximum EEG engagement index of the participant, but often the values observed during the calibration step did not reflect the values seen during the learning tasks. During the pilot studies, calibration had already been observed as a possible issue of the system, which led to the decision to change the task performed for calibration (Prinsen et al., 2022). The issue that emerged from the pilot studies was that participants seemed to have a way lower flat engagement index in the calibration, probably due to the novelty effect. In an attempt to fix this flaw, the calibration task was changed so that the participant would be interacting with the robot. Moreover, the robot was programmed to give an extensive explanation on the nature of the experiment before the calibration, in the hope to mitigate the novelty effect. Although this approach seems to have worked in some cases, the modality of selection of the minimum and maximum engagement is still far from optimal, and the calibration task used (a version of the n-back task) might have not been the best choice for this task. The non-consistent efficiency of the calibration task had, understandably, a great influence on the amount of third repetitions triggered in the adaptive condition. There were some instances in which the third repetition would be always triggered, and some in which it never happened. It is important to notice that the third repetition not being triggered is not necessarily a sign of the system not working correctly. If the participant is actively engaging throughout the whole task, a third repetition when not needed could be distracting. However, this concept can only work if the EEG engagement index collected really reflects the engagement of the participant. As shown in the results, in this experiment the adaptive condition definitely triggered the third repetitions more times than the non-adaptive one. This was to be expected since the probability of a third repetition in the non-adaptive conditions was set to be quite low. The results also showed that the correlation between amount of third repetitions and test score was not significant, meaning that the second sub-question can be answered by saying that the number of third repetitions does not necessarily influence the language score. However, considering the issues with the number of adaptations caused by the issues with calibration, this answer might not be generalizable. Further research on this aspect is still needed. It could be valuable to investigate not only the effect of the number of repetitions in different experiments, but also the impact of different types of interventions on the learning outcomes.

The issues with the adaptability of the system might also explain the reason behind some of the results of Godspeed's questionnaire on the impressions of the robot. The results showed that there was no significant difference in the comparison between adaptive and non- adaptive. For concepts such as intelligence, where intuitively an adaptive robot should score higher, the faulty adaptations of the adaptive robot might have had a great influence on the score. However, despite the issues with the system, the results from the System Usability Scale questionnaire seem quite promising in terms of how the entire system is perceived. It was not expected to see great differences in the results of this questionnaire between the two conditions and that was indeed the case. However, the participants seem to have given quite high scores overall. This means that even if the system can definitely be improved on the technical side, it is still perceived as good enough, so the overall structure of the system can be reused in different experiments. In addition, the results of the User Engagement questionnaire showed that, even though there was no significant difference in the level of self-reported engagement between conditions, the self-reported reward had a tendency to be higher for the adaptive condition. Therefore, the answer to the third sub-question is that there is no significant effect of the condition on the self-reported user engagement, but it can be interesting to investigate it further. An optimized version of the system might lead to more promising results, as a better adaptation to the EEG engagement index might lead to an improvement in the self-reported engagement for the adaptive condition.

6 CONCLUSION

In this thesis, a BCI system to monitor engagement, connected to a robot tutor, was presented. The main aim was to investigate whether a robot tutor adapting to the engagement level monitored by the EEG engagement index could improve learning. The results have shown that no direct correlation between the adaptive condition and higher test scores could be found. Overall, this experiment did not lead to the desired results, but it could be an important stepping stone for further research. The BCI system developed for this project can definitely be improved, but has already proven to be functional. In addition, the feedback from the participants showed that the system, including the robot, seems to be well received, meaning that the same setup could possibly be replicated in different experiments. Further research is needed to get a more accurate overview of the impact of an EEG-based system with a robot tutor on learning outcomes. Some important points that could be improved in the system are the calibration part, as seen in the discussion, and the way the adaptation of the robot is implemented. The calibration could benefit of an automatization of the selection of minimum and maximum engagement. This would probably already give more consistent results. Moreover, further research on which calibration task to use could be very beneficial. As for the adaptations, the approach chosen in this experiment was to make the robot say a short sentence pointing out that the participant's engagement seemed low, followed by a third repetition of the word, with the associated gesture. This is only one of the many possible approaches, and more research on the effect of different types of interventions could make a great impact on the outcomes of the experiment. Furthermore, the field of adaptive learning based on EEG data, especially when applied to robots, is definitely worth exploring further. There are still a lot of components that can be investigated and could potentially lead to important discoveries. Overall, this research can be seen as a valid starting point to know which parts might need a more in-depth analysis, and what can be done to improve the BCI system proposed.

REFERENCES

Alimardani, M., Braak, S. v. d., Jouen, A.-L., Matsunaka, R., & Hiraki,

K. (2021). Assessment of engagement and learning during child-robot interaction using eeg signals. In *International conference on social robotics* (pp. 671–682).

Bajaj, R., & Sharma, V. (2018). Smart education with artificial intelligence based determination of learning styles. *Procedia computer science*, *132*, 834–842.

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.

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

Coelli, S., Sclocco, R., Barbieri, R., Reni, G., Zucca, C., & Bianchi, A. M. (2015). Eeg-based index for engagement level monitoring during sustained attention. In *2015 37th annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 1512–1515).

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. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. doi: 10.1145/3171221.3171277

Ewing, K. C., Fairclough, S. H., & Gilleade, K. (2016). Evaluation of an adaptive game that uses eeg measures validated during the design process as inputs to a biocybernetic loop. *Frontiers in human neuroscience*, *10*, 223.

Khedher, A. B., Jraidi, I., & Frasson, C. (2019). Tracking students' mental engagement using eeg signals during an interaction with a virtual learning environment. *Journal of Intelligent Learning Systems and Applications*, *11*(01), 1–14. doi: 10.4236/jilsa.2019.111001

Mohamed, Z., Halaby, M. E., Said, T., Shawky, D., & Badawi, A. (2019). Facilitating classroom orchestration using eeg to detect the cognitive states of learners. In *International conference on advanced machine learning technologies and applications* (pp. 209–217).

Morgan, H. (2014). Maximizing student success with differentiated learning. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *87*(1), 34-38. Retrieved from https://doi.org/10.1080/00098655.2013.832130 doi: 10.1080/00098655.2013.832130

Mubin, O. (2011). *Roila: Robot interaction language* . Technische Universiteit Eindhoven.

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.

Prinsen, J., Pruss, E., Vrins, A., Ceccato, C., & Alimardani, M. (2022). A passive brain-computer interface for monitoring engagement during robot-assisted language learning. *Submitted to 2022 IEEE International Conference on Systems, Man and Cybernetic*.

Pruss, E., Prinsen, J., Vrins, A., Ceccato, C., & Alimardani, M. (2021, 12). A bci-controlled robot assistant for navigation and object manipulation in a vr smart home environment.. doi: 10.5220/0011010800003123

Rohani, D. A., & Puthusserypady, S. (2015). Bci inside a virtual reality classroom: a potential training tool for attention. *EPJ Nonlinear Biomedical Physics*, *3*, 1–14.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292.

Szafir, D., & Mutlu, B. (2012). Pay attention! designing adaptive agents that monitor and improve user engagement. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 11–20).

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.

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