CHILDREN LEARNING THROUGH TEACHING AN ERRONEOUS ROBOT

“IT’S TO, NOT TOO!”: THE IMPACT OF ROBOT ERRORS ON CHILDREN’S LEARNING IN A LEARNING-BY-TEACHING PARADIGM

BY HUNTER KENNEDY CERANIC, B.ENG & SOCIETY

A Thesis Submitted to the School of Graduate Studies in the Partial

Fulfillment of the Requirements for the Degree Master of Applied Science

April 25, 2025

McMaster University

Master of Applied Science (2025)

Hamilton, Ontario (Department of Computing and Software)

TITLE: “It’s To, Not Too!”: The Impact of Robot Errors on Children’s Learning in a Learning-by-Teaching Paradigm

AUTHOR: Hunter Kennedy Ceranic (McMaster University)

SUPERVISOR: Dr. Denise Y. Geiskkovitch

NUMBER OF PAGES: Roman Numerals, Regular pages

Lay Abstract

Access to quality education key factor for addressing societal problems, but as teacher resources are continually being spread thinner this goal becomes more difficult to achieve. Introducing educational tools, such as teaching robots has been shown to have comparable effects to being taught by human tutors, which may help alleviate the burden on teachers. There has been exploration using robots as tutees in learning-by-teaching, as the paradigm has been shown to provide better learning outcomes than standard methods. In this thesis, we investigate the design and utilization of strategic robot errors in a learning-by-teaching scenario to improve children's learning. While we did not find significant results regarding the cognitive learning efficacy of different mistakes, post-hoc analysis was performed indicating that certain mistakes impact affective characteristics that contribute to learning such as attention and self-efficacy. Implications of this research for robot design research and applications are discussed.

Abstract

Access to quality education is widely acknowledged as a key factor to address the problems we face globally as a society. Teacher resources, however, continue to dwindle due to lack of funding and increasing demands for more personalized education, making it difficult to find the time to help each student succeed. As such, the introduction of educational tools, such as social robots for extra 1-on-1 help, has been shown to have comparable effects to being taught by human tutors, which may help alleviate the burden on teachers.

In this thesis, we present an experiment investigating the design and utilization of strategic robot errors in a learning-by-teaching scenario to improve children's reading ability. The experiment tested three different conditions to help differentiate the best strategy for developing robot errors: *targeted mistakes* designed to engage the zone of proximal development, the challenge level of problem-solving for optimal learning, *simple mistakes* which are easy and obvious to identify requiring little effort on the part of the tutor, and *no mistakes* which acted as a baseline.

While we did not find significant results regarding the cognitive learning efficacy of different mistakes, post-hoc analysis was performed indicating that certain mistakes impact affective characteristics that contribute to learning such as attention and self-efficacy. The implications of this research for robot design, research and implementation and broader applications in society are discussed, as the use of mistakes to influence affective learning outcomes may be effective at overcoming other known shortfalls of the

technology. In addition, recommendations in regard to improving experimental methodology for future studies using robot tutees and future research directions for robot error design are explored.

Acknowledgments

There are so many people who contributed their knowledge, experiences and time to my education to support me writing this paper, who I want to acknowledge here. First and foremost, I would like to thank Dr. Denise Geiskkovitch, for supervising this master's thesis and everything that includes, planting that spark of curiosity about the academic side of HRI for me in my final year of undergraduate studies and entertaining my whimsical writing style.

Along with Denise, I would like to thank Dr. Irene Ye Yuan and Dr. Shane Saunderson for being members of my M.A.Sc. defence examination committee. I want to say a huge thank you to everyone in the HuRoN lab for being so friendly, sharing laughs and our passion about how robots and humans fit together. A special thanks goes to Dr. Julia Rosen for your incredible advice, and telling me about Swedish academic time, I will be using that. Another huge thank you goes to Divya Patel for being the best wizard I could ask for, and being a voice of reason when I started to worry about things outside of my control.

Finally, to my family, Victor, Michele, Logan, Ayden and Finnick Ceranic, thank you for being there to care for me and push me to chase my passions. To my partner through these past 2 years, Nyah Lawryshyn, thank you for being my side through the trials and tribulations of this thesis. I hope you see this paper as an apology for having to deal with the stress I had writing it. And, to the Engineering & Society program, thank you for molding me into who I am today – an engineer who cares.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

CLI	Command Line Interface
DRA	Developmental Reading Assessment
GRL	Guided Reading Levels
HRI	Human-Robot Interaction
SAR	Socially Assistive Robot
SDGs	United Nations Sustainable Development Goals
SMA	Simplified Miscue Analysis
WoZ	Wizard-of-Oz
ZPD	Zone of Proximal Development

Declaration of Academic Achievement

I, Hunter Kennedy Ceranic, declare that this thesis titled, ""It's To, Not Too!": The Impact of Robot Errors on Children's Learning in a Learning-by-Teaching Paradigm" and the work presented in it are my own.

1. Introduction

Human-Robot Interaction (HRI) is a multi-disciplinary research field concerning study of the relationships between humans and robots from a range of perspectives. It is a very broad field encompassing issues relating to the social and physical design of robots, the applications of robots in the wider context of society, as well as how robots influence and are influenced by culture [1]. HRI researchers investigate the applications of robots in a variety of roles including education, healthcare and therapy, entertainment, service jobs, personal assistants, search and rescue [1]. To help organize this broad field, researchers of HRI have identified distinct categories of robots, based on their design goals and capabilities. This thesis is concerned with one specific category of robot – socially assistive robots (SARs), which are designed to provide assistance to human users, through social interactions using both verbal and non-verbal communication modalities [1, 17]. SARs have been studied for use in specific roles such as medical and elderly care, coaching, and companionship [1, 17]. This paper explores the design and implementation of an SAR for applications in child education. Specifically, we designed SAR error behaviours to enhance tutee robots in the learning-by-teaching paradigm. This teaching strategy was selected in particular, as it has been found to correlate with the highest retention rates for learned material when compared to other traditional teaching methods [15, 58], therefore arguably providing the highest quality learning outcomes.

In 2015 the UN passed a motion to adopt a set of 17 sustainable development goals (SDGs) to be accomplished by the year 2030, in an effort to foster global peace and

prosperity [56]. The fourth goal on this list is *Quality Education*, which emphasizes the notion of creating inclusive and equitable education for all [56]. The UN has identified that reaching the goal of *Quality Education* is a key enabler for achieving all of the other SDGs [56]; with increased access to a higher standard of education we will have more people able to collaborate and develop ideas to solve the problems posed by the other SDGs. There is a pressing need to address this goal, as according to the UN education is trending downwards globally [19]. They project that by 2030, without proper interventions being developed, 84 million children will be out of school, and 300 million children will lack basic numeracy and literacy skills [19]. Approximately 80% of the 104 countries studied by the UN reported learning losses due to the pandemic, highlighting how delicate of a goal high quality education is [2]. Equitable access to education and increasing the resources available to teachers is especially important as low and low-middle income families are falling even further behind – there is nearly a $100 billion dollar annual financing gap for these families to reach their education goals [19]. While there are many different potential solutions to attempt to move education systems towards higher quality outcomes, one promising direction is optimizing the delivery of high-quality education through expanding and improving the tools educators have available to them. These types of solutions can help alleviate the stress on educator's finances and time. While, traditionally children are taught by adults or peers, research in HRI has explored how robots can be utilized in educational settings as a potential tool for teachers [5, 6, 36, 38, 39, 67], and the findings warrant further investigation.

The usage of tools such as virtual agents and tutoring software to deliver education has been explored for many years [5], demonstrating varying levels of success [29, 30]. These technological interventions in education have been developed to address concerns relating to limited school budgets, class sizes and the growing demand of children who need individualized help, straining teacher resources [5]. SARs have been introduced as promising alternatives to teaching agents, adding a physical aspect to the delivery of automated education. This real-world physicality, typically referred to as embodiment, also gives rise to more effective social elements in the interaction between humans and robots, such as non-verbal communication. Findings in pedagogy research have shown that social interaction between teachers and students in educational environments leads to enhanced learning outcomes [66], and it has been demonstrated that these benefits can extend to interactions between robots and humans [1, 40, 49, 52]. In comparison to virtual agents, SARs are generally perceived in a more positive way likely due to their embodiment enhancing social interactions [5, 25]. Furthermore, robots tend to be more engaging, leading to higher attention and compliance over other pedagogical technologies [5, 33]. HRI researchers continue to explore the application of SARs, especially in pre-tertiary education, due to these promising results [39]. Studies have been conducted where robots assist in teaching subjects such as language learning [9, 21, 24, 26, 32, 68], logic and pattern recognition [31, 40, 51, 52], as well as math and science [33], though the last two tend to be underrepresented in the literature likely due to the social nature of robots lending itself better to language studies [36, 38, 39, 67]. Design choices for improving learning outcomes are also a focus of research. Communicative gesturing behaviours [64, 65, 68], timing with

regards to breaks between learning tasks [42], physical appearance and perception of the robot [14, 16, 25], and social communication styles [20, 28], [26, 27] have all been investigated to discover how to best enhance robot effectiveness in educational roles. Notably, since their control systems are governed by customizable code, robots have the capability to offer highly personalized education. When robots employ teaching strategies which are tailored to individual students, this approach has been found to enhance engagement and produce cognitive and affective learning outcomes comparable to those of human counterparts [2, 5, 39]. SARs have emerged as prime platforms to research how we can utilize existing learning frameworks in an effort to deliver high quality education without further straining teacher resources.

Pedagogical studies have indicated that peer tutoring can be one of the most effective methods of learning, specifically for the tutor [15, 46, 58]. This is especially true when the tutee is trying to learn or ask questions about a topic the tutor is not an expert in but can reflect on their previous experience to help them problem solve and develop solutions to best aid the tutee [15, 46]. Due to evidence that SARs can be viewed as peers by children [14, 39], learning-by-teaching has also been explored with robots in the place of human peers [7, 11, 23, 41, 52], with findings showing promising opportunities as related to learning and pedagogy. Furthermore, robot customizability gives rise to the potential of robot tutees being the perfect tutee for tutor learning; targeted questions and mistakes, if designed correctly could arguably result in the best learning outcomes possible. For this reason, we investigated the design of robot errors in the learning-by-teaching education

paradigm to better understand the effects of a robot tutee's mistakes on children's learning outcomes.

1.1 Research Motivation

Compounded with the studies which have been performed by the UN indicating that there is a need for solutions which move society towards more accessible quality education [19]. The primary motivation for this thesis is exploring if it will be beneficial to use SARs as a tool in an effort to contribute to solving this societal problem.

Although there have been studies regarding the use of robot tutee errors in the learning-by-teaching paradigm, this is still a relatively new field of research, and as such some aspects of the use of robots in the paradigm have yet to be fully explored. Given the promise of a future where we can take full advantage of the personalization robots offer in educational roles, it is imperative to ascertain the best behaviour protocols that are conducive to learning. First and foremost, it is important to gather evidence which suggests that the mistakes robots make impact learning. The learning outcomes from a session where a child tutors a robot versus a child watching and listening to a robot complete a problem successfully should be different enough to provide a justification to use the technology as educational tools.

For younger children, fully realizing the benefits of being a tutor in learning-by-teaching interactions can be challenging due to the cognitive demands needed to internalize

and acknowledge ones own problem-solving capabilities [15]. However in certain modalities of learning-by-teaching such as correcting and giving feedback, reflection on the error being made is an automatically required component of the tutor role in order to successfully teach the tutee [15]. In this scenario, the tutor may actively problem solve at a higher level than they are usually comfortable with to identify and correct the mistake, without consciously having to acknowledge and explain why the mistake was made. In this master thesis, the goal is to determine if designing robot mistakes in a learning-by-teaching scenario to intentionally approach the this level of problem-solving, commonly referred to as the zone of proximal development (ZPD) [60], of a younger child is sufficient enough to achieve the improved learning outcomes typically associated with the paradigm.

In this thesis we will test the hypothesis that mistakes that are designed to target the general ZPD of the child tutors will always lead to the best learning outcomes. Since the learning-by-teaching paradigm focuses on learning by correction, there are circumstances where tutors may only elicit *knowledge telling* behaviours, which are summary explanations that do not engage the ZPD and lead to lesser learning outcomes [15, 46]. Therefore, we want to further examine if there is a significant difference in children's learning outcomes when robot mistakes are targeted towards the ZPD compared to mistakes which they could easily identify independently. These mistakes that would be easily identifiable for child tutors given their reading level, are more likely to induce *knowledge telling* behaviours since they are within the children's comfort zone, and they will not need to engage in breaking down the problem to find the error. This is important to test as it could point towards whether the underlying mechanisms of learning-by-teaching are not

able to be fully taken advantage of by SARs due to their weaker verbal communication capabilities [52], or if augmentation of the paradigm by using intentionally designed robot tutee mistakes is a sufficient solution to maximize the learning outcome benefits.

With the results of this study presented and discussed using standard experimental methodologies in HRI learning-by-teaching interaction design, this thesis aims to determine the impact of the quality of robot errors on learning outcomes and lay a foundation for future research on the use of robot tutees and the design of their mistake behaviours.

1.2 Research Objectives

The objective of this thesis is to evaluate the effect of robot errors on the learning outcomes of 6-8-year-old children who act as a tutor for a robot tutee during a reading task. In the experimental study we conducted, a robot read a book out loud, alongside participants who were tasked with correcting the robot if it made any reading mistakes. To evaluate our hypotheses, we used a between-participants design where each participant was randomly assigned to one of three study conditions, representing the three types of behaviours the robot was programmed to display: reading with *no mistakes* (the baseline condition), reading with *simple mistakes* (which are designed to be comfortably within their capabilities and easily identifiable), and reading with *targeted mistakes* (which are designed to approximate the tutor's zone of proximal development). In total, the valid results of 27 participants were collected. The learning outcomes of the participants were analyzed based

on the score difference of pre- and post-experiment quizzes. The experimental sessions were also recorded as references for post-hoc data collection. The following hypotheses were made to guide our research objectives:

H1) In the *targeted* and *simple mistake* conditions participants will achieve better learning outcomes, or will show more improvements across quiz scores, than the participants in the *no mistakes* condition.

H2) Participants in the *targeted mistake* condition will achieve better learning outcomes, or will show more improvements across quiz scores, than participants in the *simple mistake* condition.

2. Related Works

The purpose of this chapter is to highlight previous research, which was referenced to support the creation of the research objectives, and the development of the experiment discussed in this thesis.

2.1 Digital Agents in Education

While teachers remain a central part of education, digital agents and other similar technological tools have been used in pedagogy research in an effort to assist in teachers' delivery of education. Digital agents are programmed computer-controlled characters, which can be interacted with by students in a social manner, usually via text or images on a monitor [11]. By using knowledge of pre-existing social schemas, the designers of such agents aim to have them mimic traditional methods of learning via social interaction [7]. Digital agents in these applications have been designed to give advice, and demonstrate thought processes for learning [11, 48, 53], as well as attempt to influence more affective components of learning such as motivation through encouragement [3, 11]. The idea is that in these roles as teaching aids, digital agents can go beyond what normal people can achieve, due to an infinite amount of patience and time resources, and the ability to communicate via visualization of concepts and thought processes [7]. These advantages translate to the data where the use of these technological tools have been found to improve students learning by moderate amounts, from the 50^{th} to 75^{th} percentile, though these results

are typically skewed towards individualized tests created for research rather than standardized tests [30].

Though they have some successful results, they are typically inconsistent especially in relation to the socio-emotional outcomes of learning [29], as the use of digital agents as tools for teaching could be hindered by the abstract nature of the agent itself. Existing only on the two-dimensional plane of a computer monitor, reduces the ability for a tutor to identify and relate to it in a social capacity, as findings suggest the more human-like technology appears the easier it is for people to accept it having intelligence [41, 57]. For this reason, robots which can have a physical three-dimensional embodiment similar to a human, and are able to better display collaborative social behaviours, have a higher potential as pedagogical tools [41].

2.2 Robots in Education

Similarly to digital agents, robots can take on the role of tutors to support and supplement the educational strategies of teachers in new ways. This way of thinking is reflected in the research regarding robots in education; as of 2018, 48% of the pedagogical studies involving robots placed them in the role of a tutor [5]. In a study by Serholt et al., the overall efficacy of robot tutors compared to human peer tutors was investigated in a study consisting of 27 students aged 11–15-years-old [51]. It was found that the mere presence of a physical robot caused children to be more eager to learn, however students were more comfortable with asking questions to human tutors to due the limited verbal

capabilities of the robot [51]. According to Belpaeme et al., robot tutors have an impact on cognitive learning outcomes, which affect knowledge retention and understanding, comparable to those of human tutors [5]. Robot tutors have slightly less impact on affective learning outcomes, which are related to development of emotional or personal factors of tutees which motivate learning, in comparison to human tutors, albeit still having an effect in around half of the studies Belpaeme et al. reviewed [5]. Moreover, many related works also mention how these individual factors such as self-efficacy, attention and personality can influence both learning outcomes and study results [5, 23, 39]. While it is difficult to control for these factors, qualitative data can be collected to glean insight into what types of impacts they can have, and how we can design robot behaviour to better address them. This thesis puts an emphasis on investigating cognitive learning outcomes, however since our motivation is to provide the highest quality outcomes possible, observational data will still be collected regarding the potential impact of SAR mistake design on affective learning outcomes.

Consistently, studies involving SARs in education are constrained to specific subjects or lesson plans, and robot behaviour is often limited, with very little adaptation to the individuals learning [5]. However, there is no indication that these learning outcomes are applicable to more general tutoring applications outside of these controlled settings, and as such the technology must be developed more. We argue that the best direction to develop SAR technology towards improve learning outcomes would be to design their behaviour in alignment with evidence-based learning strategies. Accordingly, this study investigates novel robot behaviour strategies for integration into the learning-by-teaching education

paradigm, which has been found to be extremely effective for reinforcing and fostering the retention of knowledge [15, 23, 46, 58].

2.3 Learning-by-Teaching

To interpret findings from HRI studies involving learning-by-teaching, it is necessary to first explore the underlying mechanisms of the paradigm. In tutor-tutee dynamics, the goal is typically for the tutor to teach the tutee. However, research into this dynamic has revealed that the tutor also engages in learning-by-teaching, where through their instruction of the tutee, the tutor reinforces their own learning [15]. This effect is emphasized further when the dynamic is peer-to-peer instead of adult-to-child, as it enables current students to further enrich their quality of education [15, 58].

Learning-by-teaching has been shown to involve different cognitive processes when compared to learning for oneself, due to the expectancy that someone else will rely on them to teach the material. Awareness of this expectation causes the tutor to place greater effort into understanding and organizing the content in meaningful ways [15]. This is a phenomena known as the protégé effect, and arises due to the tutor's feelings of responsibility to the tutee [11, 23].

Tutors have the potential to experience higher quality learning outcomes than even their tutee, as a result of their improved task commitment in combination with their interactions explaining concepts to the tutee [15]. Roscoe and Chi argue that these

enhanced outcomes only occur when tutors engage in a form of collaboration in learning-by-teaching interactions called *reflective knowledge building* [46]. Generally, tutors, though more experienced in the subject than the tutee, are likely not experts, which means there are knowledge gaps they must account for when teaching. To do so, the tutor has to perform a very demanding task: assess their own understanding of a problem and ensure their explanations and guidance make logical sense [15, 46]. This has the additional benefit of often making the explanations of tutors who engage in this thought process, higher quality [15, 46]. This is the tutor recognizing and grappling with their own ZPD. However the problem arises that oftentimes instead of *knowledge building*, tutors tend to prefer what is called *knowledge telling* [15, 46]. Tutors engaging in *knowledge telling* mainly summarize knowledge, akin to memory recall, resulting in shallower explanations and less learning benefits for both parties [46]. Since there seems to be a latent bias towards *knowledge telling* [46], whether a tutor engages in *knowledge building* is also reliant on if they have received training to do so [15]. This makes *knowledge building* especially challenging for younger learners in primary school, due to the introspection and experience it requires, resulting in less learning benefits. For this reason it may be the case that younger children often play the role of a collaborator or tutee instead of a tutor in a majority of studies with educational robots [39, 51].

On a glance, it might seem that previous research in learning-by-teaching indicates that the benefits of the paradigm may not be able to be fully leveraged by younger child tutors. However, different modalities of learning-by-teaching are easier to engage with for younger children. For example, learning by correcting mistakes and providing feedback is

a very common form of peer learning. In these scenarios tutors inherently must reflect on their own knowledge, and break the task down to examine how their tutee deals with a problem and learn from their peers mistake [15]. The fundamentals of this paradigm are essential to understanding how the robot behaviour we describe in this thesis was designed and programmed. Part of the goal of this thesis is to design mistakes that deliberately attempt to elicit *knowledge building* behaviours, in an effort to enhance learning outcomes. Understanding learning-by-teaching also provides clarity regarding the decisions made by other researchers who have investigated using robots as tutees.

2.4 Robots as Tutees

When investigating the usage of robots in a learning-by-teaching scenario, researchers attempt to explore the benefits of using robot tutees and how their unique capabilities can be further leveraged in this role. While we have discussed some of these benefits directly, they are best highlighted by a series of studies as part of the START project spearheaded by Pareto and Serholt translated findings from a teachable agent to a robot-tutee and then continued to directly compare the use of a robot tutee versus a child tutee in a learning-by-teaching scenario [41]. In two parallel studies they had 12–13-year-olds participate in a gamified mathematics-based tutoring scenario with either a robot or a 9–10-year-old tutee [40, 52]. They collected the general perceptions tutors had of the tutee during the session, regarding how much they learned and their enjoyment tutoring, and used this qualitative data to assess the effectiveness of the interaction [40, 52]. Their results

suggest that generally robot and human tutees perform comparatively similar in the role, both evoking high enjoyment in the task, and creating an environment where the tutor believes that learning occurred [40, 52]. Verbal communication and collaboration in the interaction was where the robot's capabilities fell short, causing them to rely on outside intervention from adult teachers to help [40, 52]. However, tutors indicated they preferred to have adult teachers in the room during the experiment with both types of tutees regardless, to help guide the interaction [40, 52]. Furthermore, there was evidence that tutors felt that they did learn moderately more in the robot tutee condition [52]. This may have been due to child peers often asking fewer questions and generally eliciting less interaction than their robot tutee counterparts [40]. We know that collaboration is essential to knowledge acquisition [60], and that communication and collaboration are glaring weaknesses of robot tutees [40, 52], however acknowledging this we can attempt to strategically circumvent scenarios which rely on these capabilities. For this reason, this thesis focuses on leveraging other modes of learning-by-teaching, by designing SAR behaviour to engage tutors in *knowledge-building* without the need to rely on heavy amounts of collaboration.

2.4.1 Robot Tutee Error Behaviours

Robots are inherently prone to errors, due to the nature of the computers and the computer software that control them. HRI research has investigated the impact of robot errors on people's perception of robots. Such studies have found that, while simpler or harmless errors can actually make people feel more favourably towards the robot, and elicit

positive reactions like laughing and smiling [35, 55], more severe errors provoke escalating social responses and can lead to a decrease in trust in the robot [47, 55]. Similar results have also been found with young children, showing that verbal informational errors made by robots can negatively impact children's trust [18]. However, these studies take place outside of the context of an educational setting. When the context changes and the child's expectation is to teach the robot, the errors robots make instead lead to a greater engagement in the learning task [33].

Given these findings, research in HRI has been conducted to evaluate how to best implement these mistake behaviours in tutee robots. Hood et al. conducted a study with an emphasis on creating a learning-by-teaching scenario where mistake behaviour was perceived to be authentic (i.e., the mistakes did not seem intentional) by the tutor. In the study, they created an algorithm that enabled the robot to physically enact the motions of handwriting, and subsequently develop "poor" handwriting shapes [21]. As 6–8-year-old child tutors observed the robot write, they would correct the robot through demonstration, and the robot would respond to this, altering the shape drawn according to the feedback [21]. While the study did not explicitly test for learning outcomes, they found that every child perceived this type of learning scenario as believable, and as a result in 9 of 14 of the sessions children continued to try to teach the tutoring period, indicating that it evoked a high level of engagement [21].

Chandra et al. elaborated on these findings by using a similar handwriting scenario to measure 7–9-year-old children tutor's perceptions of how much the robot tutee learned, and analyzed if that impacted their own learning outcomes [9, 10]. They compared a robot

that continually improves on their mistakes in response to tutoring and a robot who does not improve [10]. A follow-up study added another condition for comparison where the robot copied the feedback given by the child tutors in an effort to develop personalized learning [9]. Additionally, they collected information on the participants perceived self-efficacy regarding their performance in the role as a tutor [9, 10], but interestingly not in regards to their own confidence in handwriting. In the first study findings suggest that continuous improvement, or the perception that the robot was learning from their mistakes did in fact result in significantly better learning outcomes [10]. This was however not replicated in their follow-up study, where there was no significant difference found between the learning outcomes of the personalized, continuously improving and non-improving mistake conditions [9]. Chandra et al. accounted for this discrepancy by noting their small sample size (25 participants in the first study and 37 in the second), and how other factors such as the child's attention and motivation may have impacted the results.

In contrast, a study by Yadollahi et al. focused on how robot gesture behaviours can be used to help children identify robot tutee mistakes in reading-based learning-by-teaching scenarios [68]. To accomplish this, the researchers characterized three different types of reading mistakes that 6–7-year-old children make at their reading level using the *Simplified Miscue Analysis* (SMA) framework [13, 68]. They then implemented these mistakes as robot errors, and evaluated how robot gestures, such as pointing at the word being read, impacted the children's discovery of the mistakes [68]. The authors found that, while pointing itself did not have a significant impact overall on the number of corrections the child tutor was able to make, pointing did help children identify mistakes that particularly

correlated with illustrations [68]. Again however, while the learning outcomes of the interaction were not evaluated, the researchers did comment on the children having high levels of engagement in the task [68].

Overall, while there have been some studies conducted on the impact of robot mistake behaviours in learning-by-teaching, there remains large gaps in the research with regards to the actual learning outcomes they create and the underlying mechanisms that contribute to them.

3. Experimental Methodology

In this thesis we investigate the impact of robot mistake design for use in learning-by-teaching scenarios with children. In this chapter we outline the methodology – including the participants, technology and materials we used, how the robot mistakes we tested were developed, and as well as the experimental procedure.

3.1 Participants

In accordance with previous studies conducted utilizing robot tutees in learning-by-teaching scenarios we identified the age range of 6–8-year-old children as the best fit for our study [9, 21, 68]. Thirty-one participants (M = 7.13, SD = .81, 16 girls, 15 boys) were recruited at the Hamilton Public Library, in Hamilton, Ontario, Canada, with a majority recruited at the Turner Park Branch. At the Turner Park branch, a large portion of the recruitment was performed alongside the reading buddies program for 6–12-year-old children, though children from outside the program were recruited as well. As a thank you for participation, parents of participants received $15 CAD, and their child received a small toy, regardless of if the experimental session was completed or not.

Since we were working with an underage population receiving informed consent to participate in the study was of the utmost importance. To achieve this, we established a multi-step consent process. First we developed a letter of information for parents and guardians which provided an outline of what their child would experience in the study and

any potential risks that they would incur. They were also informed that they would receive $15 CAD, and their child would receive a small toy as compensation for their participation in the study. Most importantly, we provided information about how their child's data would be collected through video recording, and allowed them to choose how much of their child's image and likeness we were allowed to use upon the dissemination of our findings. Their signatures were collected to indicate their consent. An assent script was developed to communicate the same information contained within the letter of information to the participant at their understanding level. Samples of these forms can be found in Appendix A – Ethics Forms. Parents/guardians of participants were required to provide consent before the experimental session began.

A debrief script was also developed to explain that information was withheld from the participants, specifically that they were actually the ones being tested not the robot. This choice was made to avoid introducing bias into the experimental sessions, which we explained to the participants at their understanding level. Samples of the debriefing script can also be found in Appendix A – Ethics Forms.

The study described in this thesis was submitted to the McMaster Ethics Review board and received ethics clearance as project MREB#6063.

3.2 Conditions

The participants were randomly assigned to one of three conditions to test our hypotheses – *targeted*, *simple* and *no mistakes*. This subchapter gives a brief background of the theory we used to create *targeted mistakes* and the development of both the *targeted* and *simple mistakes*.

3.2.1 The Zone of Proximal Development

In order to design and code robot mistakes to enhance the quality of education that they can deliver we must first understand how humans best learn. In *Mind in Society: The Development of Higher Psychological Processes*, Vygotsky proposed that knowledge is socially constructed [60]; it is through interaction with other people that we begin to internalize and come to "know" the things that we are taught. To explain the underlying mechanisms of how this works Vygotsky coined the term *Zone of Proximal Development* (ZPD). He defined the ZPD as the "distance between the actual developmental levels as determined by independent problem solving and the level of potential problem solving under adult guidance or in collaboration with more capable peers" [60].

Figure 1. A diagram illustrating the Zone of Proximal Development [63].

The ZPD describes the range of problem difficulty that is conducive to an individual's cognitive growth in any given learning area. These problems can only be solved through assistance and collaboration with others who have more experience or knowledge in the given domain. In order to guide the individual who is trying to learn how to solve a problem the more capable person must develop scaffolding, engaging the learner's ZPD by breaking down the problem into smaller, more manageable tasks [36, 60]. Social constructivism and the principle of scaffolding are typically used to describe what happens during interactions in one-on-one tutor-tutee scenarios. As a result, HRI researchers, working under the assumption that the impact educational robots have are comparable those of humans [5], can leverage this framework to create novel robot behaviours targeted at enhancing learning outcomes. As a tutor this may manifest as robot

behaviours such as asking targeted questions, however it also enables robots to occupy the role of a tutee and use intentional mistake behaviours in an effort to scaffold learning to target a child's ZPD [5, 36, 41].

3.2.2 Mistake Development

Creating personalized robot behaviours has the potential for improved learning outcomes, however the study by Chandra et al. could not prove definitively that this was the case [9]. Due to the large amount of individual data we would have to collect from participants about their reading level prior to conducting the study, we decided that individualized mistakes would not be appropriate given the scope of this thesis. Instead, we opted to design the mistakes in the study around the average reading level of the population's age range, which presented clear benefits and drawbacks. Given that learning-by-teaching is inherently a strategy for reinforcing what the tutor has already learned [15], designing the mistakes for the average reader instead of personalizing them will make identifying and correcting errors require less scaffolding resources [58], potentially leading to better learning outcomes. We acknowledge however that all of the children would likely start at different reading levels, though we believe this drawback could be mitigated through random assignment of participants which would account for this variance across the conditions. In addition, we took extra care in selecting the reading material to be used for the study which is further discussed in Section 3.4 Materials.

To begin designing the robot errors we started with the *targeted mistakes* condition as it proved to be the largest challenge: How do we design believable mistakes to ensure

they fall within the ZPD of children 6–8-years-old? Fortunately, we found the system for characterizing mistakes adopted by Yadollahi et al. could also be largely adapted for our purposes, as they developed their mistakes for children of the same age range [68]. To define the three categories of reading mistakes for their study Yadollahi et al. used the SMA Framework created by Cunningham in 1984 [68]. The SMA defines guidelines that can be used to analyze the root of any reading error made by a child [13]. These guidelines are organized into a series of four questions: "*Did the miscue look like the original wording*?", "*Did the miscue leave the syntax of the passage essentially the same*?", "*Did the miscue leave the meaning of the passage essentially the same*?", and "*Did the reader successfully correct the miscue?*" [13]. Since the goal of the interaction is to have the tutor correct the mistake being made, the first 3 questions were the main references used when developing mistakes.

In addition Scarborough's *Reading Rope* framework was also referenced when developing *targeted mistakes*, as it is typically used by teachers to facilitate reading comprehension in young children [62]. We identified six of the eight skills outlined in the *Reading Rope* could be incorporated into to our mistake design to ensure they were using the full range of their reading skills in the study: *Background Knowledge* referring to the context of the text, *Vocabulary* referring to the range of how words can be used, *Language Structures* referring to word syntax and semantics, *Phonological Awareness* referring to awareness of how words should sound, *Decoding* referring to understanding how phonemes combine to create a word and *Sight Recognition* referring to the automatic processing of words without decoding [50]. The other two skills *Verbal Reasoning* and *Literacy*

Knowledge [50] are more difficult to test in one reading, since they are based on inferencing and understanding literary genres and as such we were not able to incorporate the ideas these skills represented into the *targeted mistakes*. Given this set of criteria *targeted mistakes* often resulted in subtle changes from the original text such as words that have the similar morphology or pronunciation. **Error! Reference source not found.** in Appendix B – Robot Mistakes contains the list of all the *targeted mistakes* used in the robot script.

For *simple mistakes* another set of criteria had to be developed, to ensure the mistakes were easy or obvious to identify. In contrast to the *targeted mistakes* condition, the goal was to have the children engaging in *knowledge telling* when identifying *simple mistakes*. As mentioned before, the SMA represents the root causes of typical reading mistakes made by children [13], therefore we found that creating mistakes which did not correlate with the SMA criteria would make them strikingly obvious. Each of the *simple mistakes* were developed using the same criteria: the mistake would be unrelated to the original phrase regarding both syntax and semantics and would be words that are part of common vocabulary for 6–8-year-old children. A table containing the list of all the *simple mistakes* used in the robot script for the condition can be found in Appendix B – Robot Mistakes.

The final condition was the *no mistakes* condition. In this condition, the robot read the original text as normal. However, it is important to note that regardless of condition the participants were informed that the robot may make a mistake, so they all had an expectation that they would be participating in a learning-by-teaching scenario.

3.3 Technology and Setup

3.3.1 NAO

An integral part of this project was the choice of robot platform to act as a tutee. For this experiment we used Softbank Robotics' NAO, a commonly used SAR in Child-Robot Interaction research, and especially in learning-by-teaching research [9, 21, 23, 68]. The robot features the NAOqi OS which is fully programmable, and offers SDKs in a multitude of programming languages [37], of which we chose Python for comfortability and ease of programming. NAO is a humanoid robot which is 23 inches tall and has 25 degrees of freedom [37]. This proved to be beneficial for our project as we designed the interaction such that NAO would stand up and sit down, move its head between the reading and the participant, and gesture with its arms, to enhance the authenticity and child engagement in the interaction. Furthermore, NAO comes prepacked with text-to-speech capabilities [37], which we utilized to have the robot read book passages out loud and communicate with the participant about mistakes. For the purposes of the experiment, we gave the robot a gender-neutral name, SAM-EE, to help the participants recognize the robot as a peer, and to avoid introducing gendered bias.

3.3.2 Wizard-of-Oz (WoZ)

When designing the mistake generation algorithm, we decided to not to make the robot fully autonomous. The reasoning for this decision was two-fold: firstly since we were

working with 6–8-year-old children, we had to ensure that any interactions or mistakes made by the robot were age appropriate and consistent with the criteria we developed in 3.2.2 . Furthermore, by using an interaction set-up called Wizard-of-Oz (WoZ), we could control the robot if needed to generate more specific or detailed responses to make the scenario more genuine. WoZ is a very commonly used experimental technique in HRI research for enabling semi-autonomous operation of robots, to make it appear as if the robot is behaving autonomously [45]. In traditional WoZ set-ups, the robot is placed into a position to interact with the participant, and a "wizard" who is hidden or otherwise not participating in the interaction observes and controls the robot's behaviour as needed [45]. The use of WoZ streamlined our robot design, as the wizard that was controlling the robot through a command line interface (CLI), could listen to and interpret what the children were saying directly and respond accordingly, as opposed to needing voice recognition algorithms to detect and parse what they were saying. This also enabled us to essentially script the robot's behaviour removing the need for computer vision to read; NAO would simply enact a pre-determined set of movements while following the story incorporated into its code for each condition, and when necessary the wizard could manually deviate from that script, through custom typed responses at any time. The wizard was trained on how to use the CLI to control the robot, and briefed on the experimental procedure from the perspective of the robot prior to the study to ensure smooth operation when interactions with the participants occurred. The GitHub Repository located at [8], contains the modules that were used to control the robot's behaviour in the experiment.

3.4 Materials

The book *Frog and Toad are Friends* by Arnold Lobel [34] was chosen for the reading activity. This choice was made for a variety of reasons, the first, being that it addressed the concern of accommodating the skill range of 6–8-year-old readers. We consulted the Scholastic Readers Wizard to choose a book that aligned with the Guided Reading Levels (GRL) and Developmental Reading Assessment (DRA) scales which are common reading level guidelines used in schools across North America [44]. Second, the book is a collection of short stories, which meant one story could be selected to reduce the amount the participants need to read and therefore the length of an individual study session could be shorter. The story that we chose to use for the experiment was *The Letter*. Furthermore, this book is older (it was originally published in 1970), so our participants are much more likely to have not read it before.

The book chosen was also available as a PDF, which was important, as the robot "read" off a tablet in the experiment. We decided to use a tablet because the text could be made larger such that it could be placed in a position which made it believable that the child and the robot could both read from it. This positioning also made it so that children could not touch the tablet to change the page, which would disrupt the robot script. This setup is pictured in **Figure 2**. In addition, the tablet's screen could be streamed allowing the wizard to track exactly which sentence was being read in the book.

Figure 2. Experimental Setup with labels. Not pictured here are a video camera behind and to the right of the setup.

3.5 Measures

To test our hypotheses, we developed a method to measure learning outcomes. Similar to the study by Chandra et al. we developed pre- and post-tests to evaluate the child's reading skill before and after the robot interaction [9]. These tests were administered to every participant regardless of condition, and their learning outcomes, defined as the difference between their pre-test score and post-test score, were compared between-conditions. The questions were read aloud to the child, and the answers were displayed on cue cards which were randomly presented. The children were instructed to select the correct answer to the best of their abilities. To design the quiz questions, we adapted the same criteria that we used to develop mistakes in section 3.2.2 Mistake Development**,** and

composed two quizzes consisting of 10 unique multiple choice questions each, which tested the same concepts. The contents of the questions were derived from passages in the book *Frog and Toad are Friends* to ensure that they were at the same reading level as the robot task. Each quiz question had two possible answers: the correct answer, and an incorrect answer that represented a mistake similar to the ones the robot might make. 9 of the questions on each quiz had incorrect answers based on *targeted mistakes* and 1 question (question #4 on the pre-test and question #10 on the post-test) had an incorrect answer based on *simple mistakes*. It is important to note that none of the quiz answers overlapped with the mistake words the robot would say during the reading to avoid the participants relying on memory to identify the correct answer instead of reading skill. The pre- and post-test forms can be found in Appendix C – Test Forms. In addition to quizzes, data for other metrics such as the number of mistakes identified by each participant during the robot task, as well as the number of pages they asked to be re-read, were also collected and compared across condition.

3.6 Experimental Procedure

To begin parents/guardians provided consent to have their child participate in the experiment. After consent was received, each participant was randomly assigned to one of the three conditions: *targeted mistakes*, *simple mistakes*, or *no mistakes*. Participants were sat down at a table, in a private room with the experimenter sitting to the right of them who proceeded to read them the assent script. In the assent script, participants were given brief

instructions about their task and explicitly told the goal of the study was to "understand if children helping robots read improves the robot's reading level". The assent script was designed to clarify to the participants that even though they were being quizzed, the robot was the one being tested. To ensure this, the assent script explained that the participant's job was to help the robot learn by correcting its mistakes, and answering the quiz questions as correctly as possible to improve the robot in the future. It should be noted participants were explicitly not told that WoZ was used to control the robot, to reduce any preconceived bias they had towards the robot's behaviour. Once verbal assent was given, the video cameras were turned on, and the first phase of the experiment began.

In the first phase of the experiment, the participants listened as the experimenter read the short story from *Frog and Toad are Friends*, called "The Letter" aloud. In this phase, the goal was to expose the participant to the reading material so that they could mentally prepare themselves to help the robot read the same book. After the reading was completed, the experimenter then administered the 10-question pre-task quiz which gave an initial estimate of their reading skills. The experimenter then left the table to bring the robot over and begin phase two of the experiment.

In the second phase of the study, the participant was introduced to the robot. The robot was placed on the table, slightly to the left of the child, such that the child and the robot both can have a clear view of the tablet which the book is read off (**Figure 3** and **Figure 4**). The wizard then followed an introductory script that was the same for every participant. The robot, controlled by the wizard, conversed with the child and explained that it needed help reading the story. The robot also informed the participant of the rules of

the interaction, asked the participant to tell the robot to stop reading when it made a mistake so that they could correct them together, and that the participant would be asked by the robot if they wanted it to read each section over again at the end of each page. After the participant responded they were ready to read, the robot was commanded by the wizard to begin the reading the story "The Letter" according to the condition the participant was allocated.

The robot read one page at a time, and the experimenter, still sitting to the right of the participant, controlled the tablet when it was necessary to proceed to the next page. Across the entirety of the reading task, the robot made 12 mistakes in the *targeted* and *simple mistake* conditions, once per page. If the participant stopped the robot because they thought it had made a reading mistake, the robot would turn its head from the tablet towards the participant and ask the child to explain what the error was. After being told what the error was, the robot would then repeat the sentence in question with the child's correction, and ask if it changed the mistake correctly, to acknowledge it was learning from the interaction. The robot would then continue reading the page now using the corrected sentence. If the child ever turned to the experimenter to ask about the robot's performance or for any other reason, the experimenter redirected them to the robot or the task. If the child ever asked the robot a question during the reading task, the robot could respond, but would usually ask the child to continue reading. Once the book was finished, the wizard initiated the ending script which was the same for every participant. In the script the robot would thank the child for helping it, say goodbye, and then would be carried away by the experimenter to avoid distraction. The experimenter would then proceed to administer the

post-task test, which, similarly to the pre-task test, was used to evaluate if there were any improvements made through participation in the learning-by-teaching interaction.

Figure 3. The experimental setup as viewed by the right-side camera. In this photo we can see a participant talking to the robot after the robot had made a mistake. The participant is pointing to the tablet from which the book is being read.

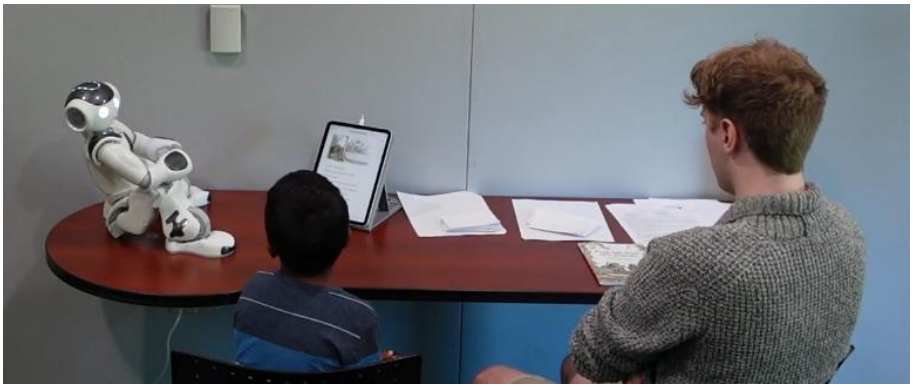

Figure 4. The experimental setup as viewed by the rear camera. In this photo we can see a participant and the robot looking at each other and interacting. The experimenter sitting to the right is supervising the interaction.

Once the post-task test was completed, the experimenter read the debrief script to the participant. The participant was given a chance to re-assent or withdraw their assent, and then they were told the experiment was complete. The participant and their parents were then given compensation, as described in section 3.1 Participants. After the experiments concluded, using the recorded videos, assenting participant's total scores were tallied and behaviours such as mistakes they pointed out, were coded.

4. Experiment Results

Out of the 31 participants who were recruited, 1 participant was removed due to lack of continuous assent, 1 due to technical difficulties, and 2 due to obvious guessing during the quiz tasks (e.g., closing their eyes and pointing at a random answer cue card). Thus, we analyzed data from 27 children aged 6-8 (M = 7.19, SD = .83, 15 girls, 12 boys), 9 assigned to the *targeted mistakes* condition, 10 to the *simple mistakes* condition, and 8 to the *no mistakes* condition. Additionally, all statistical tests discussed in this section were administered in IBM SPSS [22].

4.1 Hypothesis Results

4.1.1 Hypothesis I

To restate, the first hypothesis posited that the participants in the *targeted mistakes* and *simple mistakes* conditions would demonstrate better learning outcomes than the baseline condition. This was because in the *targeted* and *simple mistake* conditions it was expected that the participant would take on the role of a tutor in a learning-by-teaching paradigm and correct the robot's mistakes, instead of simply reading alongside the robot as expected in the *no mistakes* condition. The statistics for quiz score differences across the respective conditions are as follows: $\tilde{x}$ = 0.000, $\bar{x}$ = .778 (n = 9, IQR = 2.00, SD = 1.715) for *targeted mistakes*, $\tilde{x}$ = 0.000, $\bar{x}$ = .400 (n = 10, IQR = 1.50, SD = 1.429) for *simple mistakes*, and $\tilde{x}$ = 0.000, $\bar{x}$ = 0.000 (n = 8, IQR = 2.50, SD =

1.414) for *no mistakes.* To test this, we compared the difference between the pre- and post-test results for each condition. After performing a Shapiro-Wilk normality test it was found that there was not a normal distribution of the data. As a result, we compared these outcomes across the conditions using a non-parametric Kruskal-Wallis Test, the results of which can be seen in **Figure 5.** Kruskal-Wallis Test Results Comparing Learning Outcomes represented by Quiz Score Difference vs. Condition. We did not find any statistically significant differences in relation to learning based on score differences ($H_{(2)}$ = .688, p = .709). Therefore, based on these tests we could not reject or retain our hypothesis.

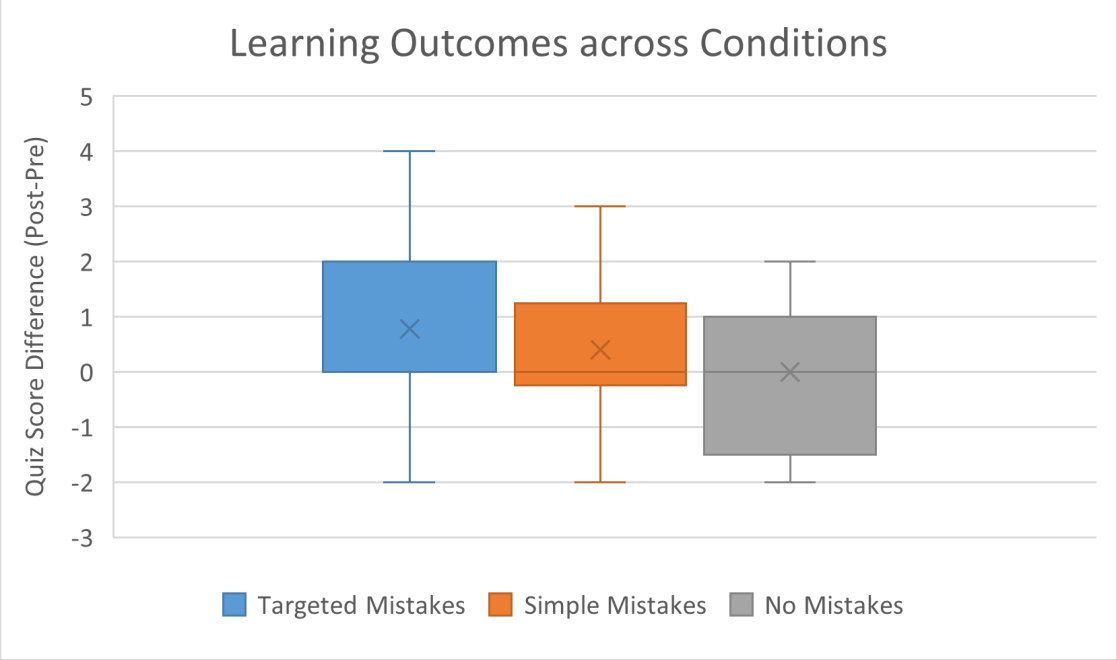

Figure 5. Kruskal-Wallis Test Results Comparing Learning Outcomes represented by Quiz Score Difference vs. Condition

4.1.2 Hypothesis II

The robot mistakes in the *targeted mistake* condition were designed to provide enough of a challenge for participants to engage their zone of proximal development, in contrast to the *simple mistake* condition where mistakes were designed to be intentionally obvious and trivial for participants to identify. We hypothesized that this would produce better learning outcomes. To test this**,** we compared score difference between pre- and post-test results for just the *targeted* and *simple* mistake conditions, whose median, interquartile range, mean and sta. A Mann-Whitney U test was performed (refer to **Error! Reference source not found.**) and it was found that there was no statistically significant difference between the score differences between the two conditions ($U = 50.5$, $p = .661$). As a result, again we could not reject our hypothesis.

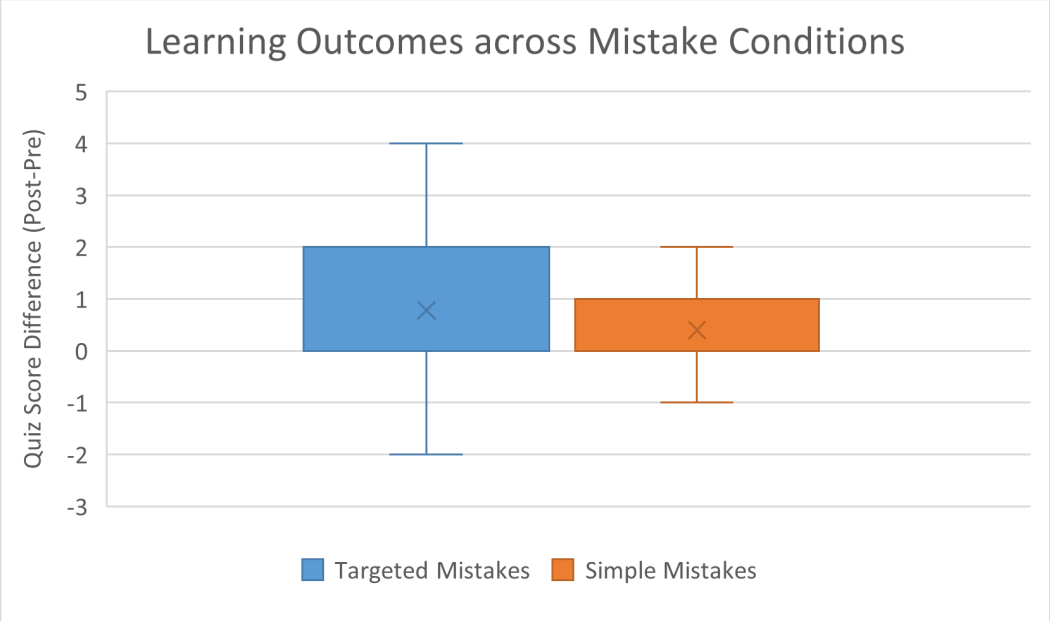

Figure 6. Mann-Whitney U Test Results Comparing Learning Outcomes represented by Quiz Score Difference vs. Condition

4.2 Post-Hoc Quantitative Results

Firstly, in order to investigate our hypotheses results, we wanted to perform a power analysis. For the statistical test we conducted for Hypothesis I, we calculated the epsilon-squared effect size of the results to be $\varepsilon^2 = 0.26$, and using that we could calculate the power of the test as $1\text{-}\beta = 0.05$ ($n = 27$). For the statistical test we conducted for Hypothesis II, we calculated the effect size of the results to be $r = 0.10$, and using that we could calculate the power of the test as $1\text{-}\beta = 0.05$ ($n = 19$). We can conclude that in both cases our results we underpowered.

Post-experiment, we performed a manipulation check to ensure that the conditions were adhered to properly. To do this, we investigated if the mistakes we designed could be differentiated by their ease of identification, which was intended when we developed the *targeted* and *simple mistakes*. We performed a non-parametric Kruskal-Wallis test comparing how many mistakes were identified by the participants across conditions, after a normality test revealed that the distribution was not normal. After performing the test, a statistically significant difference was found between all of the conditions ($H_{(2)} = 13.77$, $p = .001$), where the statistics of the conditions were as follows: $\tilde{x} = 0.000$, $\bar{x} = .778$ ($n = 9$, $IQR = 2.000$, $SD = 1.715$) for *targeted mistakes*, $\tilde{x} = 4.000$, $\bar{x} = 4.670$ ($n = 10$, $IQR = 6.000$, $SD = 4.153$) for *simple mistakes*, and $\tilde{x} = .500$, $\bar{x} = 1.750$ ($n = 8$, $IQR = 3.000$, $SD = 3.105$) for *no mistakes*. Additionally, when only comparing the *targeted* and *simple mistake* conditions the statistically significant difference held true ($H_{(2)} = 2.038$, $p = .042$), where significantly more *simple mistakes* were identified than *targeted mistakes*. This indicates

that the mistakes created did in fact result in different tutoring conditions, albeit unrelated to learning outcomes.

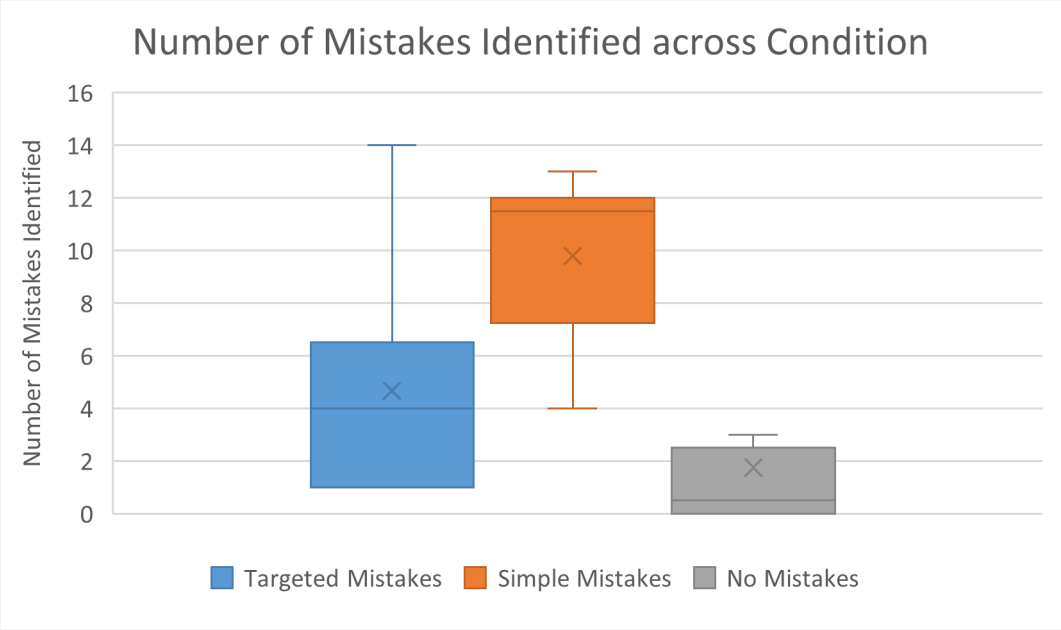

Figure 7. Kruskal-Wallis Test Results Comparing Mistakes Identified vs. Condition

We found that even in the *no mistake* condition, children still identified what they *perceived* to be mistakes where there weren't mistakes, and participated in the learning-by-teaching interaction. As a follow-up to the tests performed for4.1.1 Hypothesis I, we instead compared the number of mistakes identified to improvement, as purely engaging in the learning-by-teaching paradigm is known to improve learning [15]. To test this, we categorized participants into 3 separate groups based on their learning outcomes: *improved*, those who did better on the post-test than pre-test, *stagnant*, those who scored the same on the post-test and pre-test, and *regressed*, those who performed worse on the post-test than pre-test. The statistics for the mistakes identified across learning outcome groups are as

follows: $\tilde{x}$ = 3.000, $\bar{x}$ = 5.270 (*n* = 11, *IQR* = 11.00, *SD* = 4.962) for the *improved* group, $\tilde{x}$ = 8.000, $\bar{x}$ = 7.450 (*n* = 11, *IQR* = 8.000, *SD* = 4.655) for the *stagnant* group, and $\tilde{x}$ = 1.000, $\bar{x}$ = 2.800 (*n* = 5, *IQR* = 7.000, *SD* = 3.834) for the *regressed* group. Again, we performed a Shapiro-Wilk test and found that there was not a normal distribution of the data when comparing these learning outcome groups versus total mistakes identified. A non-parametric Kruskal-Wallis test was used to evaluate the comparison, and the results can be seen in **Figure 8.** Kruskal-Wallis Test Results Comparing Mistakes Identified vs. Learning Outcome Groups. There were no statistically significant differences found between mistakes identified and improvement ($H_{(2)}$ = 3.617, *p* = .164).

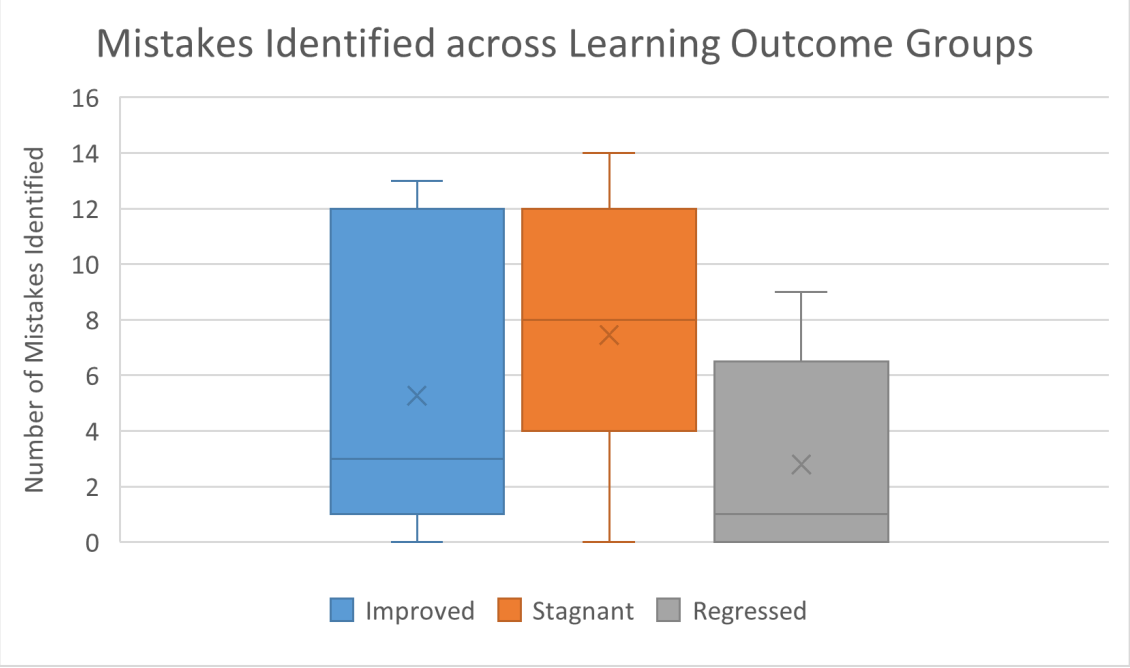

Figure 8. Kruskal-Wallis Test Results Comparing Mistakes Identified vs. Learning Outcome Groups

Since we tested with participants 6-8-years-old, we believed it would be important to investigate the impact of age on learning improvement. The distribution of the age of participants across conditions is displayed in **Figure 9** and the average age of participants sorted into each condition were as follows: *targeted mistakes* = 7.11, *simple mistakes* = 7.0, *no mistakes* = 7.5. Data trends such as the 6-year-olds having a wider dispersion of learning outcomes, 6 of 10 of the participants who maintained perfect scores being 8-years-olds, the remaining 4 being 7-years-olds, and none of the 8-year-old participants performing worse on the post-session quiz, were noted. The statistics for learning outcomes across age groups were as follows: $\tilde{x}$ = 2.000, $\bar{x}$ = .857 (n = 7, IQR = 5.000, SD = 2.478) for 6-year-olds, $\tilde{x}$ = 0.000, $\bar{x}$ = -.250 (n = 8, IQR = 2.250, SD = 1.164) for 7-year-olds, and $\tilde{x}$ = 0.000, $\bar{x}$ = .583 (n = 12, IQR = 1.000, SD = .792) for 8-year-olds. After performing a Kruskal-Wallis test comparing learning outcomes while controlling for age, there was no significant difference found between different ages and improvement ($H_{(2)}$ = 1.944, p = .378). Since we did note the wider dispersion of results for 6-year-olds, we also performed Levene's test to investigate this further and found that there is a significant difference in the variance of the age populations based on learning ($F_{(2, 24)}$ = 14.151, p = <.001). This means that, while age did not have a direct effect on learning improvement outcomes, there is a much wider variance of learning outcomes in some parts of the population than others. This aligns with the trends we identified and can be seen directly in **Figure 10**; the older participants had significantly less variance in their learning outcomes than the younger participants.

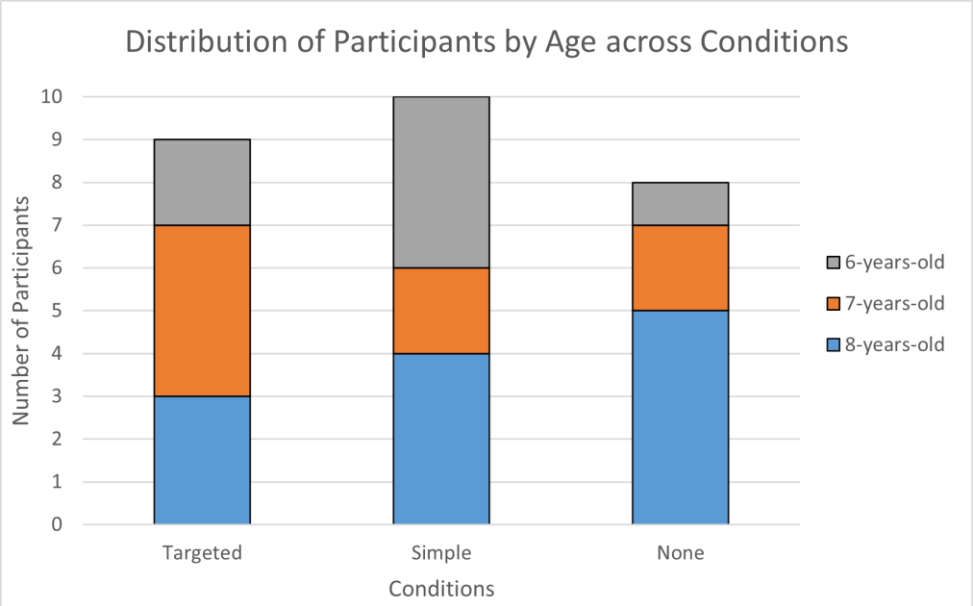

Figure 9. Composition of Conditions by Participant Age

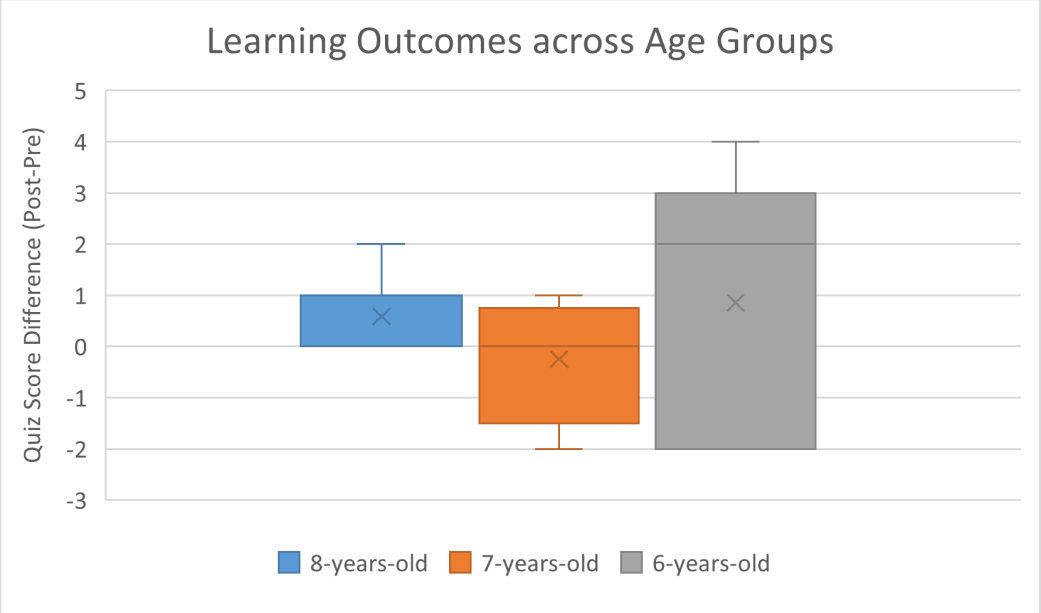

Figure 10. Kruskal-Wallis Test Results Comparing Quiz Score Difference vs. Age

4.3 Post-Hoc Coding of Behaviour Results

Observations were made during the research sessions which encouraged us to perform analysis of the behaviour participants displayed in the video data we collected. It was found that on average participants who performed better (and therefore showed improvement) on the second quiz re-read 2.55 pages (SD = 3.83, n = 11) and identified 5.27 mistakes (SD = 4.96, n = 11). Participants who scored the same on both tests re-read 1.64 pages (SD = 2.46, n = 11) and identified 7.45 mistakes (SD = 4.65, n = 11). Participants who performed worse on the second quiz re-read 1 page (SD = 1.00, n = 5) and identified 2.8 mistakes (SD = 3.83, n = 5). These slight discrepancies between performance and task participation, alongside other behaviours that were demonstrated, indicated that there was an affective component to the conditions which may have influenced the cognitive learning outcomes. This led us to develop post-hoc methods to quantitatively code the behaviour data to see if explanations for these trends would arise.

Across the conditions, the impact on attention and self-efficacy were the two main affective characteristics we were interested in investigating as they are most related to performance and task participation. To do this we developed a simple methodology to code participant behaviour in the paradigm, which we would use as checklists when watching back the videos. Two 5-item checklists were created for categories of attention and self-efficacy: the attention checklist was inspired by the ADDES-5 [54] which is a 60-item inventory used by teachers as a pre-screening for attention deficit disorder in young children, and the self-efficacy checklist was based on the BASE [4, 12] which is a 15-item

inventory used by teachers to help infer a child's "academic self-esteem". The entirety of the created checklists can be found in Appendix D – Qualitative Inventories. Each item represents a specific behaviour corresponding with either lack of attention or lack of academic self-esteem. To evaluate these checklists, we had a single coder use a simple rating scale from 0-2. The coder assigned values for each item on each checklist according to how often said behaviour was displayed by a participant. In this scale, 0 represented the behaviour never being exhibited by the participant, 1 represented the behaviour only being exhibited in one instance, and 2 represented more than one instance of the behaviour. The score assigned to each checklist item was summed per participant to give two overall *attention* and *self-efficacy* scores respectively, with higher scores representing lower levels of attention or self-efficacy. The coder in this experiment was also the writer of this thesis, who holds a B.Eng.Society in Mechatronics Engineering & Society, with a minor in Psychology. The coder received self-training on how to perform this task exactly as described above, and was instructed on how to complete this task by their supervisor Dr. Denise Y. Geiskkovitch who has over 10 years of experience performing quantitative coding.

In a similar manner to the previous quantitative tests, we then performed the Shapiro-Wilk normality test on both the attention by condition and self-efficacy by condition datasets to find that both did not have a normal distribution. Therefore, we chose to use non-parametric Kruskal-Wallis tests to evaluate the effects of condition on these characteristics. The results of these tests can be found in **Figure 11** and **Figure 13**. The statistics for rated attention scores across the respective conditions are as follows: $\tilde{x}$ =

4.000, $\bar{x}$ = 3.667 (*n* = 9, *IQR* = 3.500, *SD* = 2.449) for *targeted mistakes*, $\tilde{x}$ = 7.000, $\bar{x}$ = 6.300 (*n* = 10, *IQR* = 3.250, *SD* = 2.750) for *simple mistakes*, and $\tilde{x}$ = 5.000, $\bar{x}$ = 4.875 (*n* = 8, *IQR* = 4.250, *SD* = 2.232) for *no mistakes.* We found that there was no statistically significant impact seen on attention between the conditions ($H_{(2)}$ = 4.545, *p* = .103). However, upon performing a Mann-Whitney U Test between just the *simple* and *targeted mistake* conditions there is a statistically significant effect that occurs (*U* = 50.5, *p* = .043), showing that participants in the *targeted mistakes* condition were less inattentive than those in the *simple mistakes* condition. The statistics for rated self-efficacy scores across the respective conditions are as follows: : $\tilde{x}$ = 7.000, $\bar{x}$ = 6.222 (*n* = 9, *IQR* = 2.500, *SD* = 2.048) for *targeted mistakes*, $\tilde{x}$ = 3.000, $\bar{x}$ = 3.200 (*n* = 10, *IQR* = 2.750, *SD* = 1.988) for *simple mistakes*, and $\tilde{x}$ = 1.500, $\bar{x}$ = 2.000 (*n* = 8, *IQR* = 1.750, *SD* = 1.069) for *no mistakes.* It was also found that there was a statistically significant impact on self-efficacy between the three conditions ($H_{(2)}$ = 14.414, *p* = <.001), where participants in the *targeted mistakes* condition displayed the least self-efficacy, participants in the *no mistakes* condition displayed the most self-efficacy, and participants in the *simple mistakes* condition scored in-between the two.

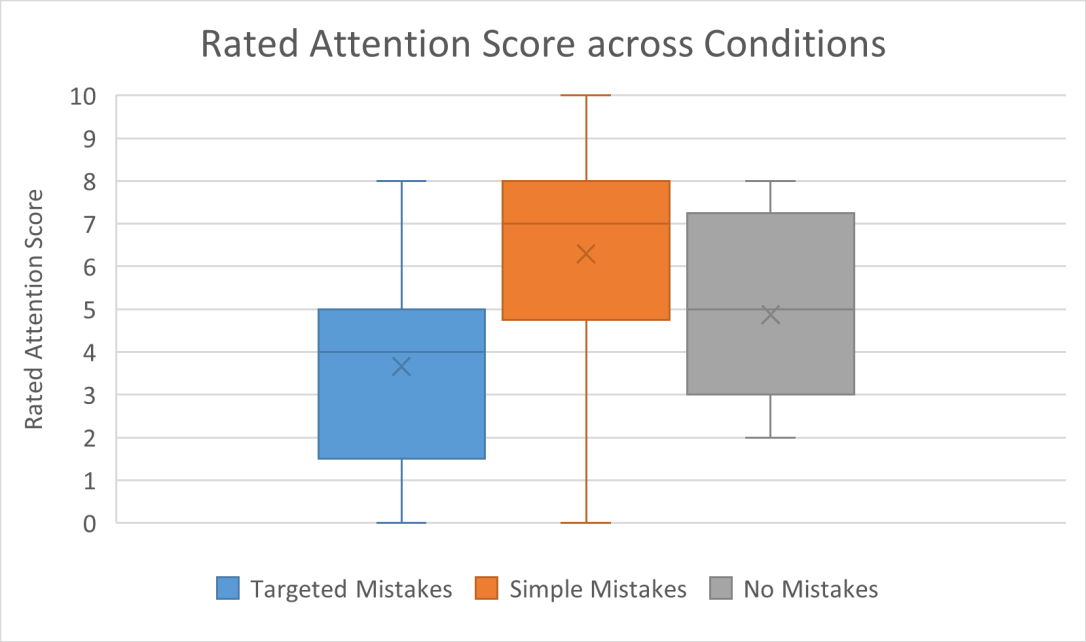

Figure 11. Kruskal-Wallis Test Results Comparing Attention vs. Condition, higher scores indicate more behaviours linked to inattention.

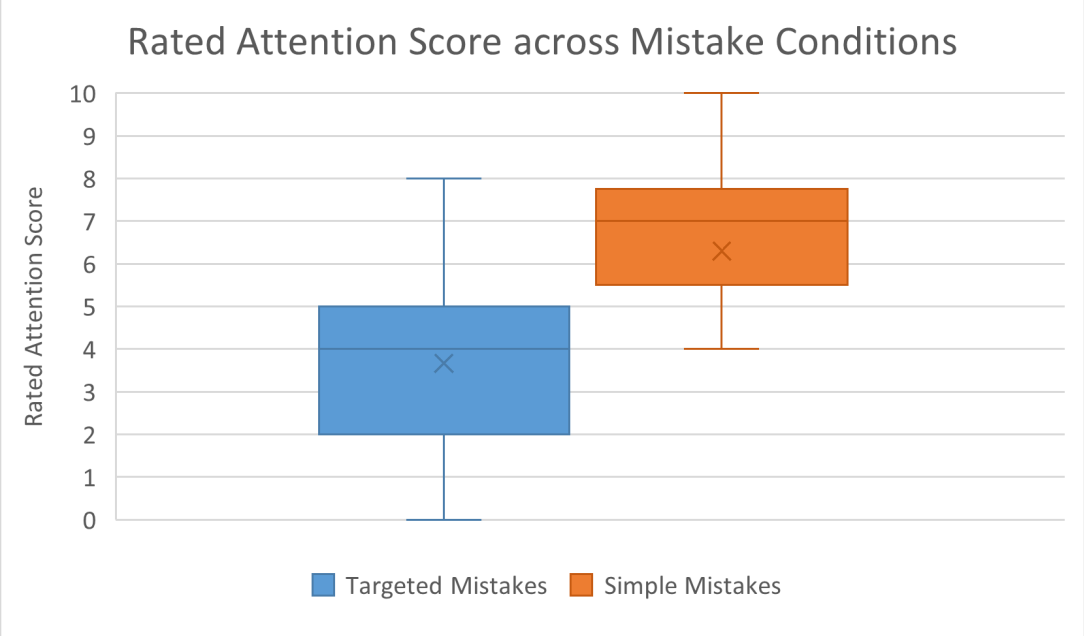

Figure 12. Mann-Whitney U Test Results Comparing Attention vs. Condition, higher scores indicate more behaviours linked to inattention.

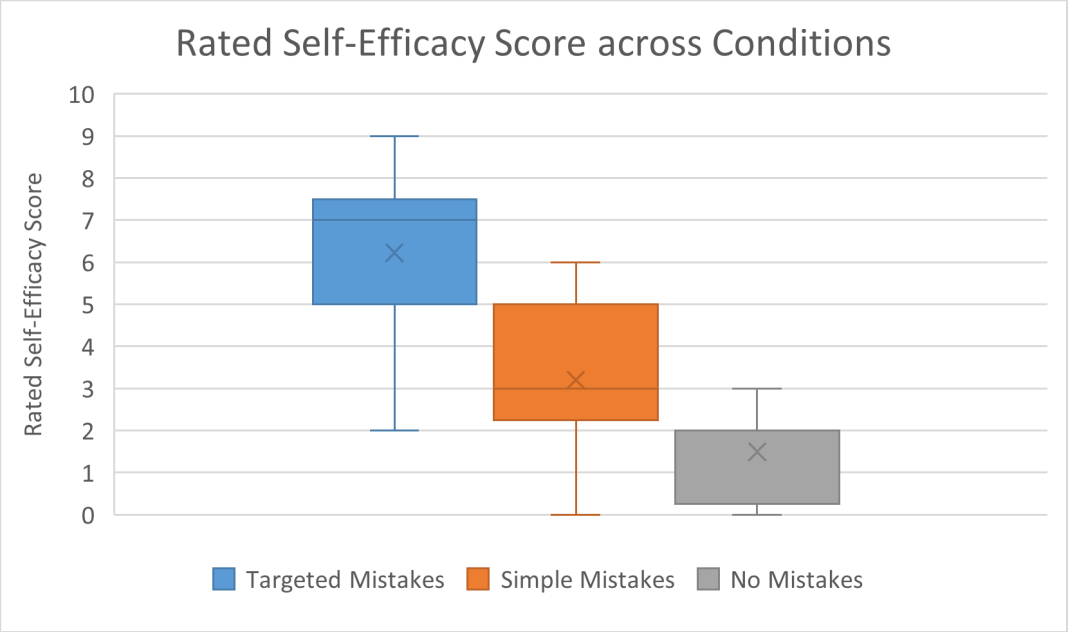

Figure 13. Kruskal-Wallis Test Results Comparing Self-Efficacy vs. Condition, higher scores indicate more behaviours linked to negative self-efficacy.

Explanations for these results and how they align with observations that were made are presented with supporting quotes from the recorded experimental sessions in section 5. Discussion.

5. Discussion

The goal of this study was to examine whether the type of error that a robot makes as a tutee during a learning-by-teaching scenario has an impact on a child tutor's learning. In this section, we look to discuss our results, and to do so we use quotes from participants who are described in Appendix E – Participant Information. Overall, while we did find that the conditions had an impact on affective factors such as attention and self-efficacy which is discussed in section 5.2 Affective Learning Outcomes, we did not find any effect of mistake-type on learning, and thus we failed to reject both our hypotheses which is discussed in section 5.3 Cognitive Learning Outcomes. Our lack of findings may have been for several reasons, and mirror prior studies which will be discussed in section 5.4 Limitations and Future Work.

5.1 The False Mistakes Phenomenon

There were some interesting observations in the *no mistake* condition that could elucidate some the results we obtained. In 4 of the 6 cases where a participant's score was maintained or there was improvement in the *no mistakes* condition, the participants, in fact, found a "mistake". These "false mistakes" were often identified due to the way the robot pronounced the words (e.g., "sent" and "send" sounding similar in the default NAO's voice), the speed at which the robot was reading making the child think there were grammar errors (e.g., One participant who said to the robot, *"Too come, you did it a little long, you*

have to make the sound "To" a little faster like "To come""(P25)), or simply due to the child mishearing what the robot said. In these scenarios, the children's behaviours were evocative of those in the *targeted mistakes* condition, likely due to the perceived challenge level of the mistakes being caught falling within their ZPD (as they were not mistakes in the first place). Otherwise, for the participants who did not hear false mistakes, the lack of mistakes led to participants exhibiting lower levels of attentiveness more akin to the *simple mistakes* condition. This duality is likely why there was no significant difference found between attention in the conditions when the *no mistakes* condition was considered alongside the others, due to this phenomenon where it could elicit behaviours similar to both. Additionally, this phenomenon may have also influenced the cognitive learning outcome results which we analyzed in our hypotheses, as it was mentioned that a majority of the participants that showed improvement or stagnation in their results in the *no mistakes* condition found false mistakes.

5.2 Affective Learning Outcomes

Post-hoc quantitative coding analysis presented results which align with key observations made during the study. We observed that there appeared to be a difference in behaviour between children who identified mistakes in the *targeted* versus in the *simple* conditions. In the *simple mistakes* condition, identifying a mistake was generally easy, which makes sense since they were designed to be obvious. We noticed that this seemed to result in improved self-efficacy for participants. As the experimental session went on, they

interacted less with the experimenter, instead promptly interrupting the robot when it made mistakes and talking over it whenever they noticed an issue. Occasionally, participants would insist the robot repeat sentences, to the point where their interactions with the robot came off as argumentative rather than teaching. For instance, one participant repeated multiple times, "*You made a mistake, you made a mistake*" (*P11*) and when the robot responded, "Please tell me my mistake", the participant said, "*Yes you made a mistake, and you did not even hear me*" (*P11*) before continuing to point out what the mistake was. Later the same participant became exasperated when the robot made a mistake saying, *"NO! It's not ball for the mail?"* (*P11*), tilting their head as if they were questioning the robot's intelligence. While much less common, this indicates that sometimes the ease of identifying mistakes may have caused children to become defensive, perhaps because they expected the mistakes to be harder to identify, resulting in a lower self-efficacy. Overall, though, this was not the case and participants in the *simple mistakes* condition displayed higher self-efficacy.

In contrast, once a child identified a mistake in the *targeted mistakes* condition, they seemed to display an increase in attention (keeping their eyes glued to the robot or the e-book), and their interactions with the robot became much more constructive. Often participants who displayed this behaviour were more cautious in ensuring that mistakes were caught, hesitating and looking for guidance from the experimenter in identifying mistakes (e.g., *"Is this the mistake?"* (*P18*)), *or* asking for the page to be re-read as they were unsure whether there was really a mistake being made. This aligns with our

quantitative coding findings which suggests participants displayed lower self-efficacy in this condition in comparison to the others.

In the *targeted mistakes* condition, attention was found to be significantly higher than in the *simple mistakes* condition. This is likely due to the nature of the *targeted mistakes* being less obvious and participants needing to pay more attention to the words being read to properly correct the robot. In the *simple mistakes* condition the mistakes were obvious, allowing the participants to pay less attention and still be able to hear the words the robot said that seemed out of place, which may align with the higher self-efficacy which was observed and displayed in the data. Conversely self-efficacy was significantly lower in the *targeted mistakes* condition than the other two conditions. This may be the result of the subtlety of the *targeted mistakes* causing participants to question whether what they heard the robot read was truly a mistake, which aligns with our observations. Additionally, if these mistakes were causing children to engage in their ZPD this could result in lower self-efficacy as well, as they might still be uncomfortable with solving those types of problems without additional assistance. These findings are further supported by previous psychological studies that suggest that as task difficulty increases, such as in the *targeted mistakes* condition, self-efficacy decreases [59] and attention increases [61].

While the attention and self-efficacy inventories do not conclusively indicate the exact impact the different mistake conditions have, they do point towards the conclusion that the mistakes a robot makes can be manipulated to influence these affective characteristics which play a role in learning. This has interesting implications for the field as typically, robots tend to have better results impacting cognitive learning rather than

affective learning [5]. Our findings indicate that intentional robot error behaviour may be a more effective strategy than the more direct verbal and non-verbal attitude reinforcement strategies that robots typically use to impact affective learning characteristics [40, 52]. That being said, both *targeted* and *simple mistakes* had positive and negative impacts, therefore a hybrid strategy that includes both of them may have to be developed to achieve the best results. Furthermore, these findings do have implications for broader society. If we can manipulate robot mistakes to influence affective characteristics, robots outside of the field of education may be able to implement similar strategies. If a robot has the potential to make a mistake, in roles such as companionship perhaps, the mistakes can be designed to impact the people around them positively. Whether that is something we would want to actually implement is an ethical consideration that should be scrutinized, but is outside the scope of this thesis.

5.3 Cognitive Learning Outcomes

Regarding why *our* study failed to reject the hypotheses we made, it is possible that the types of mistakes that we implemented, although developed based on prior literature, simply do not lead to different levels of learning. A manipulation check showed that there was a difference in the number of mistakes identified per condition, indicating the *targeted* and *simple* mistakes we designed resulted in different tutoring conditions. However, simply identifying reading mistakes may not be enough to elicit the educational benefits of the learning-by-teaching paradigm, and engage the ZPD. Multiple times, in the *targeted*

mistake condition children would look to the experimenter when they thought the robot had made a mistake, signaling they were unsure or needed help. These behaviours were also associated with comments from participants where they turned to the experimenter and asked, *"Is this the mistake?"(P18)*, *"When it says time of day it's kind of weird, it's very fast, is that like a mistake?"(P9)*, and *"Uhh I'm not sure [if the robot made a mistake], can we re-read the page?"(P15)*. These types of behaviours are consistent with what we know about the ZPD as it is the zone that children can problem solve in *with assistance* [60]. The possible explanation for this is twofold. Firstly, the *targeted mistakes* were likely designed properly and were at the level of the children's ZPD where the mistakes were difficult enough for them to recognize that it was outside the level they could comfortably identify themselves. Secondly, the participants still may have needed extra help beyond just familiarizing them with the book to develop scaffolding to properly engage their ZPD. This was an intentional limitation of the experiment design, we were directly testing to see if this was a possibility, and if these types of results are replicated it may mean that robot mistake design may not have as much of a direct impact on learning outcomes in the paradigm.

Another reason for our lack of statistically significant results may have been that the quizzes that participants took before and after the tutoring session could have been too easy, and therefore unable to properly assess learning. The reasons we believe this are threefold. Firstly, we based the quizzes on the reading itself, drawing on passages of the book to create questions. Post-experiment, we recognized this may have been a flaw in our study design, as instead of engaging in reading comprehension when answering the

questions, the participants could have used their memory of the reading to guide their answers.

Furthermore, unlike Chandra et al. [9], we decided to only have participants partake in one experimental session instead of multiple over a longer period of time. This was due to limitations regarding recruitment issues that are discussed in section 5.4 Limitations and Future Work. Again, this could have resulted in the learning outcomes measured being attributed to how much the participant learned and memorized the book that was being read, as opposed to development of actual reading skills. Finally, the results indicated the tests could have been a problem especially due to the age range of participants. For example, 8-year-olds consistently achieved 80% or higher in the pre-test, and 90% or higher in the post-test (with 11 of 12 getting 100%).

Age was not only a factor when it came to the quizzes, it may have played an overly impactful role throughout the experimental sessions. The mistakes we designed were meant to be targeting the general ZPD for children aged 6-8, and therefore was likely most appropriate for 7-year-olds. This means that the mistakes the robot made may have been too difficult for the younger participants to identify at times, and too easy for the older ones. We believe this is visible in the results, as the high variance in score difference displayed by the 6-year-olds and the low variance displayed by the 8-year-olds. More information regarding how age may have been limiting factor in the study is available in section 5.4 Limitations and Future Work

5.4 Limitations and Future Work

The study contained in this thesis had a number of limitations. For one, the sample size was quite small, especially after 4 participants had to be removed resulting in the study being underpowered. Previous experiments using robots as tutees in a learning-by-teaching paradigm have also reported similar issues regarding sample sizes [9, 21, 40, 52, 68]. In our experiment the limitation on sample size was due to difficulties with recruitment, as although we were recruiting at times the librarians recommended, there was a limited number of recruitment times we were allowed to schedule and limited numbers of children who frequented the library in that age range during those times. In addition, the length of the experiment was approximately 20-30 minutes which meant in one 4-hour period the library allocated to us, a maximum of 8 participants could be tested. Furthermore, we attempted recruiting at other libraries where we had little success with parents bringing their children to follow-up experiments. Finally, we attempted to recruit from schools, but were denied approval for various reasons such as the amount of time the participants would have to spend out of class due to the experiment. Results of the experiment may differ with a larger sample size which we recommend investigating.

Moreover, several methodological decisions were made that likely affected our findings. Firstly, the age range that we used for our sample population may have been too wide. We decided on a gap of two years to help increase potential sample size, but this impacted the design of other aspects of the experiment such as range of difficulty of the mistakes targeting the generalized ZPD. This led to the wide variance in learning outcome

results per age that was reflected in our findings. In the future, if the generalized ZPD approach is used limiting the age range to a one-year gap may be preferable. Future research could also attempt to tackle this issue by testing and adapting to each child's ZPD, as opposed to grouping children, therefore removing the need to generalize the ZPD level across an age range. If this approach is taken, the experimental methodology must be changed to include some form of pre-screening for participants, to create mistakes that adapt to their understanding level. Those mistakes also must be created on the spot for each individual participant. We wanted to avoid the need for multiple experimental sessions and the extra mistake creation preparation time when conducting our study, which is why this approach was not taken. Additionally, ensuring that the pre- and post-session quizzes are difficult enough to properly assess learning should be tested with the target population prior to conducting the experiment. In our research, we chose not to do this because the quiz questions were developed based on the SMA [13], similar to the mistakes used in the study by Yadollahi et al., [68], and due to the limited availability of recruitment times we were allotted from the library.

The presence of the experimenter during the study sessions could have also potentially affected results as participants may have been more likely to seek guidance from them, interfering with the regular flow of the learning-by-teaching paradigm. In our study the experimenter was there to help guide the interaction, ensure none of the technical equipment was damaged, and expedite the speed of transitions between the phases of the experiment. In addition, prior research has indicated the presence of an adult in the room helps facilitate learning-by-teaching interactions with young children [40, 52]. That being

said, in the future, changing the experiment design such that the experimenter is not directly beside or in the room with the participant during the robot reading phase could help address any potential interference this limitation introduced. Finally, this experiment was controlled in nature, so it is possible that these results will not generalize to in-the-wild contexts, where robots and children are engaging in improvised learning-by-teaching scenarios. Future research could conduct a similar study with fewer constraints, such as including a range of subjects being taught or a range of locations to explore this.

Beyond replicating a similar experiment to the one presented in this thesis, there are also other experiments which test a robot's efficacy in a learning-by-teaching scenario that we recommend investigating. Based on our observational findings and post-hoc qualitative analysis, future research should include additional measures for participants, such as self-efficacy, attention, and even personality measures, when looking at how robot errors may impact learning. Chandra et al. remarked that a potential reason that they were unable to replicate their own results between their two studies was that a child's attention and engagement within the scenario may play a role in their learning outcomes [9]. When we are designing robot tutee mistake behaviours, it is important to recognize, we are also essentially devising ways to test a child's mental capabilities in their current state. Testing such aspects of participants' state and personal attributes can enable researchers to better understand how other factors, such as individual differences, may influence a learning session with a robot. Interestingly, successful studies have been performed where robots as tutors suggest different studying strategies to improve motivation, and therefore learning outcomes [43], so designing interactions or mistake behaviours for robot tutees may be

similarly successful. Specifically, measuring the effect of mistakes on academic self-efficacy of children throughout learning-by-teaching sessions could shed some light into how they approach subsequent mistakes and therefore how we should design mistakes in the future to improve learning. If we can manipulate the types of mistakes that a robot makes such as using *simple* and *targeted mistakes* in tandem, to target how much a child tutor pays attention during a task, and how confident they are in their own ability to perform the problem-solving required to teach the task, we can bolster their academic learning outcomes in a more indirect manner.

Future research could attempt to investigate the phenomenon that occurred in the *no mistakes* condition, where participants found false mistakes, due to the limited vocal quality of the robot. While this problem could be addressed by using different robots or a different generated voice, further research regarding how voice quality impacts the perception of errors may be worthwhile considering it is a common area of weakness for robot capabilities [52]. Observational and qualitative data indicate that these types of mistakes may elicit behaviour similar to regular designed mistakes. Further investigating whether false robot mistakes actually lead to similar learning outcomes as true mistakes may have impacts on the field of HRI outside of learning-by-teaching.

Moreover, given that we failed to reject our hypotheses, we should acknowledge that when testing our hypotheses we were only testing if proper cognitive scaffolding could be developed by a tutor when confronted with a problem in their ZPD based on content they are familiar with. However, there are other mechanisms of the paradigm usually associated with working in the ZPD not included in our methodology. For instance

communication and collaboration in a learning-by-teaching scenario, such as tutee's asking questions about the subject matter or asking the tutor to explain their corrections, are also factors which help tutors develop scaffolding towards the ZPD [15]. Further experimentation regarding mistake design may want to put an emphasis on making room for these types of interactions in their procedure, as opposed to just simple corrections, to see if learning outcomes could be augmented by them.

6. Conclusion

Findings in the field of HRI suggest that SARs are a technology that have the potential to be leveraged for use in education, in large part due to their customizability and embodiment creating educational outcomes on par with traditional teaching. Furthermore, the learning-by-teaching paradigm, in which a child solidifies their learning through the act of teaching a less experienced peer, has been explored with SARs in the tutee role, due to pedagogical findings which indicate the paradigm is more effective at reinforcing learning than standard methods.

In this thesis we discussed an experiment we devised to test the efficacy robot tutee mistakes on learning outcomes in the learning-by-teaching paradigm. We tested three conditions, one used *targeted mistakes* toward a generalized zone of proximal development, another used *simple mistakes* which were meant to be obvious and easy for children to identify and the last was a baseline containing *no mistakes*. We conducted an experiment in which children completed a reading activity with a robot, and randomly assigned participants into the three aforementioned conditions. While we did not find an effect of type of error on children's learning, we did observe, and subsequently perform post-hoc tests which indicate that there were differences in children's attention and self-efficacy depending on the condition.

This thesis and its findings point to the opportunities that SARs can provide in the context of learning-by-teaching, as well as challenges that need to be overcome for success. Since we were unable to come to any conclusions regarding the direct impact of the types

of mistakes discussed in this thesis on learning outcomes more research needs to be performed, to replicate and further the results of this experiment. Mistake design may not have an impact on cognitive learning outcomes, and if that is the case, more emphasis should be put on developing other aspects of interactions with robot tutees in the future. However, if the results of future experiments indicate that mistake design has an impact, this would validate the choices made in previous studies with robot tutees using mistakes targeted towards the ZPD [9, 21, 68], and solidify a design goal for educational robots. Optimizing mistake design does not push SARs into a place where they will be able to solve the problem of quality education anytime soon – much more work is needed to improve their capabilities especially regarding communication clarity which was exemplified by the difficulties some participants had understanding the robot's speech. However, the results regarding the impact of robot mistakes on attention and self-efficacy are exciting for HRI researchers, and deserves further investigation in future research as it has implications outside the field of HRI as well. Evidence that intentional robot errors can be used to impact personal characteristics such as attention and self-efficacy have plenty of implications regarding robots working alongside humans in many fields not just education. When the expectation that a robot may make a mistake is present, robot designers could potentially design intentional mistakes to elicit certain behaviours, hopefully for the better, such as improving traits such as attention and self-efficacy. But as researchers, we should also be involved in the conversations about whether that is ethically appropriate to essentially manipulate people towards different mental states.

As mentioned many times in this thesis, robots are prone to errors. Robots as tutees have displayed great potential as tools for education in previous studies and in this thesis our results further point towards the idea that their erroneous behaviour can be actively leveraged to create indirect impacts on attention and self-efficacy which contribute to the act of learning. This thesis serves as a reminder to robot designers that learning has both cognitive and affective components, and that influencing one may impact the other and vice versa. While our findings regarding the impact of mistakes on the affective learning factors attention and self-efficacy have spawned strong motivations for future research, we were originally trying to impact cognitive learning outcomes, and failed in doing so. Future research must have a wholistic view of learning, incorporating both aspects, or figuring out ways to ensure they isolate them, in order to properly assess whether tutee robots can truly provide higher quality learning outcomes.

7. References

[1] Bartneck, C. et al. 2020. *Human-Robot Interaction — An Introduction*. Cambridge University Press.

[2] Baxter, P. et al. 2017. Robot education peers in a situated primary school study: Personalisation promotes child learning. *PLOS ONE*. 12, 5 (May 2017), e0178126. DOI:https://doi.org/10.1371/journal.pone.0178126.

[3] Baylor, A.L. and Kim, Y. 2005. Simulating Instructional Roles through Pedagogical Agents. *International Journal of Artificial Intelligence in Education*. 15, 2 (May 2005), 95–115. DOI:https://doi.org/10.3233/IRG-2005-15(2)02.

[4] Beauchamp, K.D. 1995. The Behavioral Academic Self-Esteem Scale with preschoolers. *Psychological Reports*. 76, 1 (Feb. 1995), 273–274. DOI:https://doi.org/10.2466/pr0.1995.76.1.273.

[5] Belpaeme, T. et al. 2018. Social robots for education: A review. *Science Robotics*. 3, 21 (Aug. 2018), eaat5954. DOI:https://doi.org/10.1126/scirobotics.aat5954.

[6] Benitti, F. 2012. Exploring the educational potential of robotics in schools: A systematic review. *Computers & Education*. 58, (Apr. 2012), 978–988. DOI:https://doi.org/10.1016/j.compedu.2011.10.006.

[7] Biswas, G. et al. 2005. Learning By Teaching: A New Agent Paradigm For Educational Software. *Applied Artificial Intelligence*. 19, (Mar. 2005), 363–392. DOI:https://doi.org/10.1080/08839510590910200.

[8] Ceranic, H. 2025. cer-hunter/NAOME.

[9] Chandra, S. et al. 2020. Children Teach Handwriting to a Social Robot with Different Learning Competencies. *International Journal of Social Robotics*. 12, 3 (Jul. 2020), 721–748. DOI:https://doi.org/10.1007/s12369-019-00589-w.

[10] Chandra, S. et al. 2018. Do Children Perceive Whether a Robotic Peer is Learning or Not? *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, Feb. 2018), 41–49.

[11] Chase, C.C. et al. 2009. Teachable Agents and the Protege Effect: Increasing the Effort towards Learning. *Journal of Science Education and Technology*. 18, 4 (Aug. 2009), 334–352. DOI:https://doi.org/10.1007/s10956-009-9180-4.

[12] Coopersmith, S. and Gilberts, R. 1982. Behavioral Academic Self-Esteem: A Rating Scale. *Consulting Psychologists Press*. (1982).

[13] Cunningham, J. 1984. A Simplified Miscue Analysis for Classroom and Clinic. *Reading Horizons: A Journal of Literacy and Language Arts*. 24, 2 (Jan. 1984).

[14] Díaz, M. et al. 2011. Building up child-robot relationship for therapeutic purposes: From initial attraction towards long-term social engagement. *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)* (Mar. 2011), 927–932.

[15] Duran, D. 2017. Learning-by-teaching. Evidence and implications as a pedagogical mechanism. *Innovations in Education and Teaching International*. 54, 5 (Sep. 2017), 476–484. DOI:https://doi.org/10.1080/14703297.2016.1156011.

[16] Edwards, A. et al. 2016. Robots in the classroom: Differences in students' perceptions of credibility and learning between "teacher as robot" and "robot as teacher." *Computers in Human Behavior*. 65, (Dec. 2016), 627–634. DOI:https://doi.org/10.1016/j.chb.2016.06.005.

[17] Feil-Seifer, D. and Mataric, M.J. 2005. Defining socially assistive robotics. *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* (Jun. 2005), 465–468.

[18] Geiskkovitch, D.Y. et al. 2019. What? That's Not a Chair!: How Robot Informational Errors Affect Children's Trust Towards Robots. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Mar. 2019), 48–56.

[19] Goal 4 | Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all | Department of Economic and Social Affairs: 2025. *https://sdgs.un.org/goals/goal4*. Accessed: 2025-03-04.

[20] Gordon, G. et al. 2015. Can Children Catch Curiosity from a Social Robot? *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, Mar. 2015), 91–98.

[21] Hood, D. et al. 2015. When Children Teach a Robot to Write: An Autonomous Teachable Humanoid Which Uses Simulated Handwriting. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, Mar. 2015), 83–90.

[22] IBM SPSS Software: *https://www.ibm.com/spss*. Accessed: 2025-02-20.

[23] Jamet, F. et al. 2018. Learning by Teaching with Humanoid Robot: A New Powerful Experimental Tool to Improve Children's Learning Ability. *Journal of Robotics*. 2018, (Nov. 2018), e4578762. DOI:https://doi.org/10.1155/2018/4578762.

[24] Kanda, T. et al. 2004. Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Human–Computer Interaction*. 19, 1–2 (Jun. 2004), 61–84. DOI:https://doi.org/10.1080/07370024.2004.9667340.

[25] Kennedy, J. et al. 2014. Comparing Robot Embodiments in a Guided Discovery Learning Interaction with Children. *International Journal of Social Robotics*. 7, (Dec. 2014). DOI:https://doi.org/10.1007/s12369-014-0277-4.

[26] Kennedy, J. et al. 2016. Social robot tutoring for child second language learning. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Mar. 2016), 231–238.

[27] Kennedy, J. et al. 2015. The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, Mar. 2015), 67–74.

[28] Kory Westlund, J.M. et al. 2017. Flat vs. Expressive Storytelling: Young Children's Learning and Retention of a Social Robot's Narrative. *Frontiers in Human Neuroscience*. 11, (Jun. 2017). DOI:https://doi.org/10.3389/fnhum.2017.00295.

[29] Krämer, N.C. and Bente, G. 2010. Personalizing e-learning. The social effects of pedagogical agents. *Educational Psychology Review*. 22, 1 (2010), 71–87. DOI:https://doi.org/10.1007/s10648-010-9123-x.

[30] Kulik, J.A. and Fletcher, J.D. 2016. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research*. 86, 1 (Mar. 2016), 42–78. DOI:https://doi.org/10.3102/0034654315581420.
[31] Leite, I. et al. 2011. Social Robots in Learning Environments : a Case Study of an Empathic Chess Companion. *Proceedings of the International Workshop on Personalization Approaches in Learning Environments* (2011).
[32] Lemaignan, S. et al. 2016. Learning by Teaching a Robot: The Case of Handwriting. *IEEE Robotics & Automation Magazine*. 23, 2 (Jun. 2016), 56–66. DOI:https://doi.org/10.1109/MRA.2016.2546700.
[33] Lindberg, M. et al. 2017. Does a Robot Tutee Increase Children's Engagement in a Learning-by-Teaching Situation? *Intelligent Virtual Agents* (Cham, 2017), 243–246.
[34] Lobel, A. 1970. *Frog and Toad are Friends*. HarperCollins.
[35] Mirnig, N. et al. 2017. To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot. *Frontiers in Robotics and AI*. 4, (May 2017). DOI:https://doi.org/10.3389/frobt.2017.00021.
[36] Mubin, O. et al. 2013. A review of the applicability of robots in education. *Technology for Education and Learning*. 1, (Jun. 2013). DOI:https://doi.org/10.2316/Journal.209.2013.1.209-0015.
[37] NAO: Personal Robot Teaching Assistant | SoftBank Robotics America: *https://us.softbankrobotics.com/nao*. Accessed: 2025-02-04.
[38] Pai, R.Y. et al. 2024. Effectiveness of social robots as a tutoring and learning companion: a bibliometric analysis. *Cogent Business & Management*. 11, 1 (Dec. 2024), 2299075. DOI:https://doi.org/10.1080/23311975.2023.2299075.
[39] Papadopoulos, I. et al. 2020. A systematic review of the literature regarding socially assistive robots in pre-tertiary education. *Computers & Education*. 155, (Oct. 2020), 103924. DOI:https://doi.org/10.1016/j.compedu.2020.103924.
[40] Pareto, L. et al. 2022. Children's learning-by-teaching with a social robot versus a younger child: Comparing interactions and tutoring styles. *Frontiers in Robotics and AI*. 9, (Oct. 2022). DOI:https://doi.org/10.3389/frobt.2022.875704.
[41] Pareto, L. 2017. Robot as Tutee. *Robotics in Education* (Cham, 2017), 271–277.
[42] Ramachandran, A. et al. 2017. Give Me a Break! Personalized Timing Strategies to Promote Learning in Robot-Child Tutoring. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (New York, NY, USA, Mar. 2017), 146–155.
[43] Ramachandran, A. et al. 2019. Toward Effective Robot--Child Tutoring: Internal Motivation, Behavioral Intervention, and Learning Outcomes. *ACM Trans. Interact. Intell. Syst.* 9, 1 (Feb. 2019), 2:1-2:23. DOI:https://doi.org/10.1145/3213768.
[44] Reading Level Scales: *https://clubs.scholastic.com/on/demandware.store/Sites-rco-us-Site/default/Product-ReadingLevels*. Accessed: 2025-01-28.
[45] Riek, L.D. 2012. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *J. Hum.-Robot Interact.* 1, 1 (Jul. 2012), 119–136. DOI:https://doi.org/10.5898/JHRI.1.1.Riek.
[46] Roscoe, R.D. and Chi, M.T.H. 2007. Understanding Tutor Learning: Knowledge-Building and Knowledge-Telling in Peer Tutors' Explanations and Questions.

Review of Educational Research. 77, 4 (Dec. 2007), 534–574. DOI:https://doi.org/10.3102/0034654307309920.

[47] Rossi, A. et al. 2017. How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario. *Social Robotics* (Cham, 2017), 42–52.

[48] Ryokai, K. et al. 2003. Virtual Peers as Partners in Storytelling and Literacy Learning. *Journal of Computer Assisted Learning*. 19, (Jun. 2003), 195–208. DOI:https://doi.org/10.1046/j.0266-4909.2003.00020.x.

[49] Saerbeck, M. et al. 2010. Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2010), 1613–1622.

[50] Scarborough, H.S. 2017. *Handbook of Early Literacy Research, Volume 1*. Guilford Publications.

[51] Serholt, S. et al. 2014. Comparing a humanoid tutor to a human tutor delivering an instructional task to children. *2014 IEEE-RAS International Conference on Humanoid Robots* (Nov. 2014), 1134–1141.

[52] Serholt, S. et al. 2022. Comparing a Robot Tutee to a Human Tutee in a Learning-By-Teaching Scenario with Children. *Frontiers in Robotics and AI*. 9, (2022).

[53] Shimoda, T.A. et al. 2002. Student goal orientation in learning inquiry skills with modifiable software advisors. *Science Education*. 86, 2 (2002), 244–263. DOI:https://doi.org/10.1002/sce.10003.

[54] Stephen B. McCarney and N. House, S. 2019. *Attention Deficit Disorder Evaluation Scale - Fifth Edition (ADDES-5)*. Hawthorne Educational Services, Inc.

[55] Stiber, M. and Huang, C.-M. 2020. Not All Errors Are Created Equal: Exploring Human Responses to Robot Errors with Varying Severity | Companion Publication of the 2020 International Conference on Multimodal Interaction. *ICMI '20 Companion: Companion Publication of the 2020 International Conference on Multimodal Interaction* (Dec. 2020), 97–101.

[56] THE 17 GOALS | Sustainable Development: *https://sdgs.un.org/goals*. Accessed: 2025-01-28.

[57] Timms, M.J. 2016. Letting Artificial Intelligence in Education Out of the Box: Educational Cobots and Smart Classrooms. *International Journal of Artificial Intelligence in Education*. 26, 2 (Jun. 2016), 701–712. DOI:https://doi.org/10.1007/s40593-016-0095-y.

[58] Topping, K.J. 1996. The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *Higher Education*. 32, 3 (Oct. 1996), 321–345. DOI:https://doi.org/10.1007/BF00138870.

[59] Voodla, A. and Uusberg, A. 2021. Do performance-monitoring related cortical potentials mediate fluency and difficulty effects on decision confidence? *Neuropsychologia*. 155, (May 2021), 107822. DOI:https://doi.org/10.1016/j.neuropsychologia.2021.107822.

[60] Vygotsky, L.S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge : Harvard University Press.

[61] Washburn, D.A. and Putney, R.T. 2001. Attention and Task Difficulty: When Is Performance Facilitated? *Learning and Motivation*. 32, 1 (Feb. 2001), 36–47. DOI:https://doi.org/10.1006/lmot.2000.1065.

[62] What Is Scarborough's Reading Rope? (Plus How Teachers Use It): *https://www.weareteachers.com/scarboroughs-rope/*. Accessed: 2025-02-03.

[63] What is the Zone of Proximal Development? 2020. *https://www.letsgolearn.com/reading-assessment/what-is-the-zone-of-proximal-development/*. Accessed: 2025-01-31.

[64] de Wit, J. et al. 2021. Designing and Evaluating Iconic Gestures for Child-Robot Second Language Learning. *Interacting with Computers*. 33, 6 (Jul. 2021), 596–626. DOI:https://doi.org/10.1093/iwc/iwac013.

[65] de Wit, J. et al. 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* (New York, NY, USA, Feb. 2018), 50–58.

[66] Witt, P.L. et al. 2004. A meta-analytical review of the relationship between teacher immediacy and student learning. *Communication Monographs*. 71, 2 (Jun. 2004), 184–207. DOI:https://doi.org/10.1080/036452042000228054.

[67] Woo, H. et al. 2021. The use of social robots in classrooms: A review of field-based studies. *Educational Research Review*. 33, (Jun. 2021), 100388. DOI:https://doi.org/10.1016/j.edurev.2021.100388.

[68] Yadollahi, E. et al. 2018. When deictic gestures in a robot can harm child-robot collaboration. *Proceedings of the 17th ACM Conference on Interaction Design and Children* (New York, NY, USA, Jun. 2018), 195–206.

Appendix

Appendix A – Ethics Forms

The Letter of Information/Consent, Assent Script and Debrief Script are found on the following pages.

McMaster University

LETTER OF INFORMATION / CONSENT

Robot Design for Young Children's Education

Student Investigator:
Hunter Ceranic
Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada
E-mail:
ceranich@mcmaster.ca

Faculty Supervisor:
Dr. Denise Geiskkovitch
Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada
E-mail:
geiskkod@mcmaster.ca

Research Assistant:
Divya Dolly Patel
Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada
E-mail:
pateld60@mcmaster.ca

Purpose of the Study:
We are collecting information on the impact of robot mistakes on child learning, in learning-by-teaching scenarios with robots. This research is part of a master's thesis project under the supervision of Dr. Denise Geiskkovitch.

What will happen during the study?
We will introduce your child to the study, as a study to "improve robot learning", and ask for their verbal assent (2 min). Your child will be read a short story by the researcher, from the book "*Frog and Toad Are Friends*" by Arnold Lobel (5 min). They will then be asked a few short questions, to test their reading comprehension based on the book (5 min). Then, they will be introduced to a child-sized humanoid robot where they will assist the robot in reading "*Frog and Toad Are Friends*", and with your permission have this session be recorded (15 min). The robot may ask your child if it has made any mistakes, so they can be corrected. Your child will then complete the same reading comprehension quiz about the book a second time (5 min). Finally, we will debrief, and your child will get a chance to select a toy as reimbursement (2 min).

Are there any risks to doing this study?
The risks involved in participating in this study are minimal. Your child may feel uncomfortable with interacting with the robot and may stop the study at any time if they find this to be the case. Additionally, if your child puts their hands near the robot's joints and the robot moves there is a potential for pinching. This will be explained to your child before they interact with the robot and the researcher will try to prevent this.

Are there any benefits to doing this study?
Your child will go through an educational learning-by-teaching exercise, which can be beneficial for their reading comprehension. Additionally, we hope to learn more about robot design in learning-by-teaching scenarios such that it can be helpful in improving educational curriculum in the future.

Incentive/Payment or Reimbursement
Parents/guardians will receive a reimbursement of $15 and children will be able to choose a small toy to take home.

Who will know what my child said or did in the study?
Your child is participating in this study confidentially. We will not use your child's name, and you may consent to the use of video of your child along with audio/video distortion for anonymity. Only the research team will know whether your child was in the study unless you choose to consent to the use of video recordings or choose to disclose this information yourself, and any identifying information will be redacted, and recordings will be deleted after the study is completed.

What if I change my mind about my child being in the study?

Your child's participation in this study is voluntary. If you decide your child will be part of the study, you or your child can withdraw from the experiment for whatever reason, even after giving consent or part-way through the study or up until March 3rd, 2025, when I expect to be submitting my thesis. If you decide to withdraw your child, there will be no consequences to you. In cases of withdrawal during the experiment session, any data you or your child have provided will be destroyed unless you indicate otherwise. If your child does not want to answer some of the questions they do not have to, but they can still be in the study. In cases of withdrawal after the experiment session audio recordings will be anonymized and unidentifiable and therefore cannot be withdrawn, however any video recordings of your child will be deleted unless you indicate otherwise. You will still receive incentive payment if you choose to withdraw.

How do I find out what was learned in this study?
We expect to have this study completed by approximately April 2025. If you would like a brief summary of the results, please provide your email address _____.

Questions about the Study: If you have questions or need more information about the study itself, please contact Hunter Ceranic at ceranich@mcmaster.ca.

This study has been reviewed by the McMaster Research Ethics Board and received ethics clearance under project 6063. If you have concerns or questions about your rights as a participant or about the way the study is conducted, please contact:

McMaster Research Ethics Office
Telephone: (905) 525-9140 ext. 23142

E-mail: mreb@mcmaster.ca

CONSENT

We will be videotaping the study session. The video will allow us to review the study session in detail, and therefore assist our data analysis. The videos will be used for anonymized research analysis. We may use anonymized quotes for purposes of dissemination; your child's name will not be included or in any other way associated with the data presented in the results of this study. You have the option of whether any video footage collected during the study session may also be used for dissemination of research results. If you do not consent, the video of your child will only be used for internal data analysis purposes. Please initial your response below:

I DO NOT consent to the public presentation of my child's participation video _____

I DO consent to the public presentation of my child's participation video:

without any modification of his/her face or voice _____
only if their face is blurred _____
only if their voice is altered _____

Please sign below to indicate that you have read and understood the information provided on this form and allow your child to participate in this study:

Name of Participant's Parent/Guardian (Printed) _____

Signature: _____ Date: _____

If you wish to receive the results of this study please record your email here: _____

Robot Design for Young Children's Education

Assent to Participate in a Study - Script

Hi, my name is **Hunter Ceranic** and I am a Student at McMaster University. Your parents/caregiver have allowed me to talk to you about a project that I am working on for McMaster University. The project is on **kids teaching robots to read**. I am going to spend a few minutes telling you about our project, and then I am going to ask you if you are interested in taking part in the project.

Why are we doing this study?
We want to find out if children teaching robots will help them learn better.

What will happen to you if you are in the study?
If you decide to take part in this study, there are some different things we will ask you to do.

First, we will read the book "Frog and Toad are Friends" together and I will ask you some questions about it
Second, I will ask you to help the robot read the story and tell them if they make any mistakes.
Lastly, I will ask you a few more questions about the story.

While doing these things all you have to do is try your best. If you have tried your best and do not know what to say or do next, you can guess or say, 'I do not know'. It will take you about 1 hour to do these things.

Are there good things and bad things about the study?
What we find in this study will be used to help improve robots. As far as we know, being in this study will not hurt you and it will not make you feel bad.

Will you have to answer all questions and do everything you are asked to do?

If we ask you questions that you do not want to answer, then tell us you do not want to answers those questions. If we ask you to do things you do not want to do then tell us that you do not want to do them.

Do you have to be in the study?
You do not have to be in the study. No one will get angry or upset with you if you don't want to do this. Just tell us if you don't want to be in the study. And remember, if you decide to be in the study but later you change your mind, then you can tell us you do not want to be in the study anymore.

Do you have any questions?
You can ask questions at any time. You can ask now or you can ask later. You can talk to me or you can talk to someone else at any time during the study.

Do you want to be part of this study?

Record of Verbal Assent:

Date: _____

Signature of the person explaining consent: _____

If assent is not given, "Thank you for your time".

Debrief and Re-Assent - Script

Robot Design for Young Children's Education

Student Investigator:
Hunter Ceranic
Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada
E-mail: ceranich@mcmaster.ca

Faculty Supervisor:
Dr. Denise Geiskkovitch
Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada
E-mail: geiskkod@mcmaster.ca

We didn't tell you the whole story about the study:
As researchers, sometimes we don't tell you everything about the study right away. This is helps us see what would actually happen in real life. Now, I will explain why we didn't tell you everything at first, and what we didn't tell you.

This study is to see how well *you* learn from the robot's mistakes. We told you, that you were helping the robot learn. If you knew the real reason, you might try harder to find the robot's mistakes, and that would change the study. So, we had to make sure you thought we weren't testing you.

We didn't tell you we were checking how much you learned from the robot. The robot is not learning from you. It is just fixing its mistakes when you tell it to.

We tried to keep secrets to a minimum, as little as possible, and hope you understand why we did this. Now we want to ask if we can keep your video data after telling you these new things. If you say no, your video and answers won't be used in the study. There is no problem if you say no.

Again, if you or your parents have any questions about the study you can ask me now or please have your parents email us using the email address we gave them.

RE-ASSENT

Do you still want to be part of this study?

Record of Verbal Re-Assent:

Date: _____

Signature of the person explaining consent: _____

If assent is not given, "Thank you for your time"

Appendix B – Robot Mistakes

Table 1. Errors used in Targeted Mistakes condition compared with the original text in the reading. The mistakes and the words they replaced are in bold.

Original Passage	**Passage updated with *Targeted Mistake***	**Mistake Development Criteria**
Frog **came** along, and said what is the matter Toad?	Frog **come** along, and said what is the matter Toad?	Sounds like original wording, leaves passage syntax and semantics essentially the same
This is my sad **time** of day	This is my sad **dime** of day	Sounds like original wording, leaves passage syntax essentially the same
No, never, **said** Toad	No, never, **says** Toad	Leaves passage semantics essentially the same
There is something that I **must** do	There is something that I **just** do	Leaves passage syntax essentially the same
He put the **paper** in an envelope	He put the **letter** in an envelope	Leaves passage semantics essentially the same
I think you **should** get up, and wait for the mail some more	I think you **could** get up, and wait for the mail some more	Leaves passage syntax and semantics essentially the same
No, **no**, said Toad	**None**, said Toad	Leaves passage syntax essentially the same
But Toad, said Frog, **someone** may send you a letter today	But Toad, said Frog, **something** may send you a letter today	Leaves passage syntax and semantics essentially the same
But there will not be **any**, said Toad	But there will not be **many**, said Toad	Leaves passage syntax and semantics essentially the same
What did you **write** in the letter?	What did you **white** in the letter?	Leaves passage syntax essentially the same
Oh, said Toad, that **makes** a very good letter	Oh, said Toad, that **maked** a very good letter	Sounds like original wording, leaves passage syntax and semantics essentially the same
Toad's **house**, and gave him the letter from Frog	Toad's **home**, and gave him the letter from Frog	Leaves passage semantics essentially the same

Table 2. Errors used in Simple Mistakes condition compared with the original text in the reading. The mistakes and the words they replaced are in bold.

Original Passage	**Passage updated with *Simple Mistake***
Frog came along and said, what is the **matter** Toad?	Frog came along and said, what is the **grape** Toad?
This is my **sad** time of day	This is my **rock** time of day
No one has ever sent me a **letter**	No one has ever sent me a **plant**
Frog **hurried** home	Frog **funny** home
to Toad's house and put it in his **mailbox**	to Toad's house and put it in his **robot**
Toad was in bed, taking a **nap**.	Toad was in bed, taking a **phone**.
The snail was not there **yet**	The snail was not there **soon**
Don't be silly, **said** Toad	Don't be silly, **match** Toad
Because now, I am **waiting** for the mail, said Frog	Because now, I am **ball** for the mail, said Frog
Dear Toad, I am **glad** that you are my best friend	Dear Toad, I am **wind** that you are my best friend
onto the front **porch** to wait for the mail	onto the front **angry** to wait for the mail
Frog and Toad waited a long **time**	Frog and Toad waited a long **book**

Appendix C – Test Forms

PRE-STUDY VERBAL TEST

1. **Toad was _____ on his front porch**. What word fits best:
    a. Sitting
    b. Feeling
2. In the story, Toad says he is, "**unhappy**". **Unhappy** means Toad is:
    a. Happy
    b. Sad
3. Frog asks about Toad getting sent mail. Frog asks, "**Not ever**?" and Toad says "**No, ____**". What word fits best:
    a. Ever
    b. Never
4. Toad says, **"Everyday my _____ is empty**". What word fits best:
    a. Mailbox
    b. Together
5. In the story it says, "**On the envelope he** _____". What word fits best:
    a. Write
    b. Wrote
6. Snail says, **"Sure, right away"**. Does this mean:
    a. Snail will do it
    b. Snail will not do it
7. Frog says, **"I have ____ you a letter"**. What word fits best:
    a. Send
    b. Sent
8. Toad says, **"I do not think _____ will ever send me a letter".** What word fits best:
    a. Anything
    b. Anyone
9. Frog and Toad go out to **"the front _____"**. What word fits best:
    a. Porch
    b. Torch
10. **____ days later the snail got to Toad's house**. What word fits best:
    a. Tour
    b. Four

POST-STUDY VERBAL TEST

1. Toad says, **"This is my sad ____ of day"**. What word fits best:
    a. Dime
    b. Time
2. Toad says, **"Everyday my mailbox is empty"**. **Empty** means there is:
    a. No mail
    b. Mail
3. In the story **Frog _____ home**. What word fits best:
    a. Hurried
    b. Letter
4. Frog, **"_____ a pencil and a piece of paper."** What word fits best:
    a. Found
    b. Find
5. **Frog ___ out of his house**. What word fits best:
    a. Run
    b. Ran
6. Frog asks snail to **"___ it in his mailbox".** What word fits best:
    a. Put
    b. But
7. **Toad was in bed, taking a ___.** What word fits best:
    a. Snap
    b. Nap
8. Toad says, **"Don't be _____"**. What word fits best:
    a. Silly
    b. Really
9. The letter says, **"I am glad you are my best friend". Glad** means Frog is:
    a. Happy
    b. Sad
10. **"Toad was very _____ to have it."** What word fits best:
    a. Pleased
    b. Envelope

Appendix D – Qualitative Inventories

ATTITUDE INVENTORY

Based on ADDES-5 Item #2

1) Is easily distracted by other activities in the environment (the wizard, movement outside the door, etc.)

Based on ADDES-5 Item #7

2) Needs oral questions and directions frequently repeated (eg., needs constant reminders to stay on task, etc.)

Based on ADDES-5 Item #37

3) Talks out of turn during the task (eg., bringing up a topic unrelated to reading, interrupts the robot to talk about something unrelated etc.)

Based on ADDES-5 Item #38

4) Moves about while seated, fidgets, squirms, etc.

Based on ADDES-5 Item #60

5) Engages in nervous habits or unnecessary movements (eg. Bites fingernails, twirls hair, chews inside of cheek, chews on objects, spins or twirls objects etc.)

SELF-EFFICACY INVENTORY

Based on BASE Item #2

1) Hesitates to make corrections or decisions (eg., participant says "I don't understand," or asks experimenter questions to confirm if a mistake was made, or clearly changes quiz answer due to being unsure)

Based on BASE Item #3

2) Does not show task independence, needs direction (eg., experimenter needs to reiterate the task multiple times to get child to engage in the task)

Based on BASE Item #4

3) Does not initiate new decisions easily (eg., requires multiple re-reads or looks for support from the experimenter to confirm a mistake, or go to the next page)

Based on BASE Item #9

4) Does not show an attitude of cooperation with the robot (eg., is confrontational or aggressive with the robot, makes threats etc.)

Based on BASE Item #15

5) Does not readily express what they perceive to be the mistakes made (eg., does not refer to the mistake word directly or uses minimal words to convey the correction to the robot)

Appendix E – Participant Information

Table 3. The age of participants and condition they were randomly assigned to are displayed in this table correlating to their participant identification number (ID). Participants who were disqualified have been omitted from this table.

Participant ID	Participant Condition	Participant Age
P1	Targeted Mistakes	7
P2	Simple Mistakes	8
P3	No Mistakes	8
P4	Targeted Mistakes	7
P5	Simple Mistakes	6
P6	No Mistakes	7
P7	Targeted Mistakes	7
P8	Simple Mistakes	6
P9	Targeted Mistakes	8
P10	No Mistakes	8
P11	Simple Mistakes	6
P12	Targeted Mistakes	6
P13	Simple Mistakes	8
P14	No Mistakes	6
P15	Targeted Mistakes	8

P16	No Mistakes	8
P17	Simple Mistakes	6
P18	Targeted Mistakes	7
P19	Simple Mistakes	8
P20	No Mistakes	8
P21	Simple Mistakes	8
P22	No Mistakes	7
P23	Targeted Mistakes	8
P24	Simple Mistakes	7
P25	No Mistakes	8
P26	Targeted Mistakes	6
P27	Simple Mistakes	7