Abstract

Creating Personalized Robot Tutors That Adapt to The Needs of Individual Students

Daniel Noah Leyzberg

2014

This dissertation makes three contributions to the study of personalization in robot tutoring: (1) we provide evidence for improved student learning gains associated with the physical presence of a robot tutor, (2) we deliver experimentally-derived design guidelines for future work in robot tutoring, and (3) we provide novel robot tutoring personalization systems and demonstrate that these systems improve student learning outcomes over non-personalized systems by 1.2 to 2.0 standard deviations, corresponding to learning gains in the 88th to 98th percentile.

We begin by investigating a foundational question in the field of robot tutoring: can the physical presence of a robot tutor affect student learning outcomes? We conducted an experiment comparing student learning outcomes between three conditions in which participants received tutoring from either: (1) a physically-embodied robot tutor, (2) an on-screen tutor, or (3) a voice-only tutor. We found that students who received tutoring from the physically-embodied robot tutor were more engaged in the lessons than students

in the other two conditions. We also found that, despite the instructional content being the same across all three conditions, students who received tutoring from the physically-embodied robot achieved significantly better learning outcomes than students in the other two groups by 0.3 standard deviations, corresponding to gains in the 62nd percentile.

In order to arrive at design guidelines for our work in automated personalization for robot tutoring, we first studied how humans personalize their tutoring. To do this, we asked participants to teach robot students, which, unlike human students, can be expected to behave in the exact same way on multiple occasions and with different human tutors. By employing robots as students, we were able to study the nuances of human tutoring personalization. We found that human tutors teach more and produce more strongly affective vocalizations to students who are less successful than to students who are more successful. We also found that, even if two students perform exactly the same on all learning tasks, human tutors still personalize their instruction based on the affective content of students' responses. We use these findings to propose guidelines for future work in automated personalization, with the goal of producing more human-like automated tutoring.

Our final contributions are our automated personalization systems for robot tutors: two of which are intended for shorter-term robot tutoring interactions and one of which is intended for longer-term interactions. For the shorter-term models, designed for use in at most one contiguous session with a robot tutor, we created an additive model intended to investigate the effects of the simplest forms of personalization systems, and a Bayesian model that is slightly more sophisticated and leads to improved learning gains over the

additive model. For the longer-term system, we used a Hidden Markov Model (HMM) that tracked students over the course of five sessions, taking place over two weeks. We evaluated these systems against similar non-personalized systems with human students and found that our personalization systems increased learning gains by between 1.2 and 2.0 standard deviations over non-personalized systems, corresponding to gains in the 88th to 98th percentiles.

Creating Personalized Robot Tutors That Adapt to The Needs of Individual Students

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

Doctor of Philosophy

by

Daniel Noah Leyzberg

Dissertation Director:

Brian Scassellati

December 2014

Acknowledgements

Of all of my supporters in this 7 year journey, I am most grateful to my Ph.D. advisor, Brian Scassellati also known as "Scaz." Scaz provided me with the opportunity to succeed here at Yale. Without his direction, none of this work would have been possible. I would also like to thank my other committee members for challenging me and strengthening my work immeasurably: Professors Dana Angluin and Drew McDermott.

I am forever indebted to my beloved parents, who have been with me through every step of the way. I celebrated every victory with them and lamented every defeat. I am an only-child and they spoiled me from the beginning with their unconditional love and undivided attention. I would not be who I am today without their guidance and support.

I am also incredibly grateful to the many friends I've met on this long journey, some of whom I am fortunate enough to also call my colleagues. Eli Kim and Lily Zeng are friends that will always hold a special place in my heart; as my closest friends, they helped guide me through many difficult decisions and moments. They supported me when I needed it the most.

I would also like to thank my many labmates and conference-mates, all of whom helped me challenge my assumptions and refine my ideas: Henny Admoni, Wilma Bainbridge, Emily Bernier, Cindy Bethel, Christopher Crick, Corina Grigore, Álvaro Castro González, Justin Hart, Brad Hayes, Alex Litoiu, Aditi Ramachandran, Fred Shick, Elaine Short, Sam Spaulding, Mariya Toneva, Danny Ullman, and Jean Zheng.

Lastly, I would also like to thank the Computer Science Education community for welcoming me as they have, first as a graduate student interloper at their conferences and now as a teaching-track faculty member. I couldn't be more excited to start my full-time teaching career at Princeton University.

Contents

List of Figures

List of Tables

"Tell me and I forget,
teach me and I may remember,
involve me and I learn."

– Benjamin Franklin

Chapter 1

Introduction

Robots that serve as social interaction partners for students outside the classroom may soon be able to provide individualized, long-term, in-home academic support that supplements a teacher's classroom instruction. Especially for students who have fallen behind in class or those who regularly need extra review and attention, such robots could serve as an important secondary source of individualized support. We envision robot tutors that function as in-home homework helpers, interacting with students one-on-one as they do their work and providing them with motivational support and content assistance. In this dissertation, we will explore some initial implementation questions of such robot tutors. Though we are not the first group to study robot tutoring, we are the first to investigate robot tutoring personalization, such that the robot personalizes the lessons it gives based on the needs of individual students.

1.1 Tutoring

In education research, one-on-one tutoring by a content expert is widely considered to be one of the most efficacious teaching modalities (**Bloom 1984**; **Cohen, Kulik and Kulik 1982**; **VanLehn 2011**). In a landmark study, **Bloom (1984)** found that students who received individual domain-expert tutoring outperformed students who received classroom instruction by two standard deviations on average, i.e. achieving scores comparable to the 98th percentile of students who received traditional classroom instruction. This result is cited as "Bloom's two-sigma effect," and it is often credited with establishing one-on-one tutoring as a gold standard against which the effectiveness of other teaching modalities and practices are measured (**Hogan and Pressley 1997**).

Though now more commonly referred to as "Bloom's two-sigma effect," this result was first called "Bloom's two-sigma problem" by the author because it highlights the relative ineffectiveness of the typical classroom instruction model that most schools and educational programs are based on today (**Bloom 1984**). Followup research has clarified that, taking into account the background of the tutor and what standards the tutor can set for his or her pupils, the benefit of human one-on-one tutoring over classroom instruction may be closer to 0.8 sigma, or test scores in the 80th percentile (**VanLehn 2011**). However, whether tutoring improves scores by 0.8 sigma or 2.0 sigma, there is a well-established positive influence of one-on-one tutors which demonstrates that our current educational system, based on undifferentiated group instruction, is producing significantly sub-optimal learning gains for its students. Perhaps, one day, with the addition of personalized robot tutors to supplement traditional classroom instruction, we can fill this learning gap and produce outcomes on par with one-on-one tutoring. Perhaps we could even produce outcomes better than typical one-on-one tutoring if we leveraged the strengths of both modalities and designed their uses to complement one another.

1.2 Human Tutoring

In order to build automated tutors that someday approach the level of success attained by expert human tutors, we must ask, "What do the best expert human tutors do that makes them so effective?" Though the mechanisms and processes are still of some debate in education research, there is a general consensus that tutors provide two functions that a typical classroom teachers do not. First, tutors give individualized scaffolded guidance to students as they solve problems or analyze new concepts by providing enough instructional support to bridge a student's knowledge gap and then iteratively taking pieces of support away until students are able to build the bridge themselves (Wood, Bruner and Ross 1976). Second, tutors gauge a student's understanding on an individual basis and build a mental model of a student's comprehension which they use to frame future scaffolding episodes (Chi et al. 2001).

In addition to acting as a safety net for students to explore the boundaries of their knowledge, tutors can also act as a significant source of motivation and accountability for students. The one-on-one interactivity of a tutoring dialogue can keep students more actively engaged in the act of problem solving and critical thinking than classroom instruction or working alone (Merrill et al. 1992). For example, prompting students to describe aloud what they've learned forces students to question their assumptions, form better synthesized conclusions, and, ultimately, increases their learning gains and retention (Chi et al. 1994; Pressley et al. 1992). More recently, Chi et al. (2001) isolated the variable of interactivity by comparing a typical tutoring interaction with a static text control group consisting of the same instructional content, finding that students new to the subject scored better on post-tests simply as a result of the interactivity of tutoring, likely as a result of increased engagement. This dissertation explores how student engagement in robot tutoring affects learning gains in Chapter 4.

1.3 Automated Tutoring

The goal of automated tutoring is to produce systems that leverage the benefits of the one-on-one teaching modality described above without requiring as many human resources. Most such systems in development today are called Intelligent Tutoring Systems (abbreviated "ITS's"). A wide variety of ITS's exist, from those designed for early childhood education for a student's first years in school (Prentzas 2013), all the way up to professional training for medical doctors (Suebnukarn and Haddawy 2004) and military personnel (Steele-Johnson and Hyde 1997). Though these systems have been in development for the past fifty years, only in the past ten years have any become commercially available. Two such commercial systems have already reached millions of students (Desmarais and Baker 2012). See Pane et al. (2014) for an account of how a commercial automated tutoring system performed in a randomized pair-matched controlled study

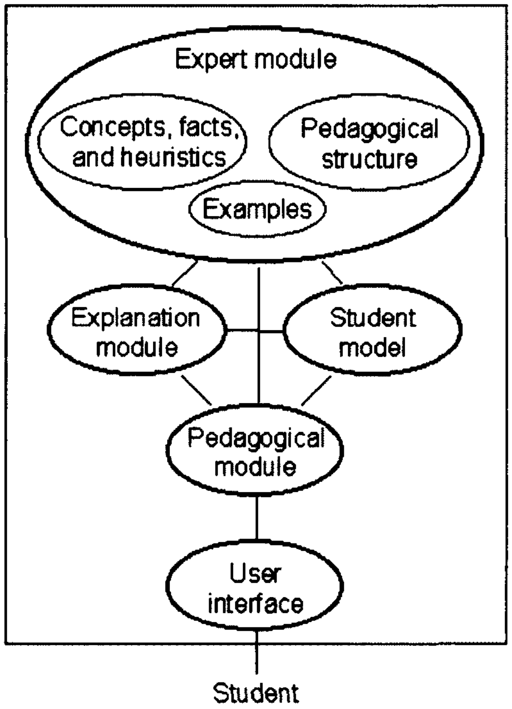


FIGURE 1.1: System architecture of FLUTE, an example of an Intelligent Tutoring System (ITS) which, like most ITS's, separates the student model from the curriculum model, seen here at the top of the diagram (**Devedzic and Debenham 1998**). Most ITS's are specialized towards the teaching requirements of a specific subject. In this case, FLUTE tutors the systems curriculum in computer science and requires content experts in that area to write its curriculum model.

with a sample size of over 20,000 students in a two year long intervention calling for supplementary use in traditional public school classrooms.

Generally speaking, ITS's are designed with four main components: (1) a student model, which tracks the progress of individual students, (2) a knowledge model, which is typically authored separately by a curriculum expert, usually a teacher, (3) a tutoring model, which is closely associated with the student model and matches available curriculum to

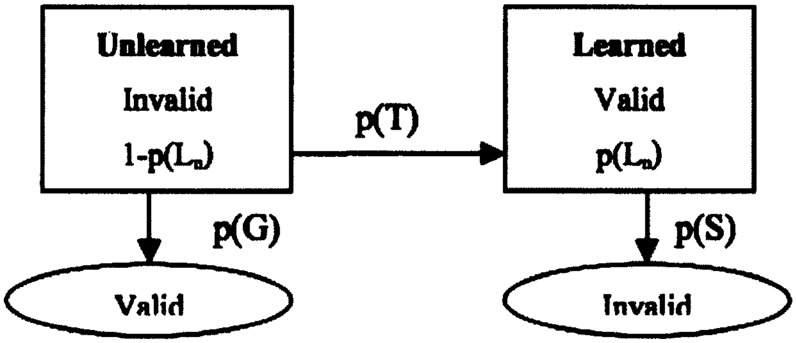


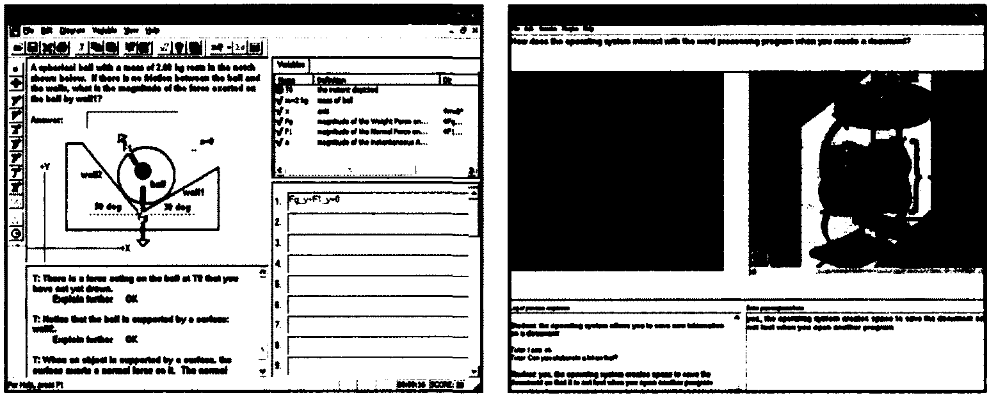
FIGURE 1.2: The Bayesian Knowledge Tracing algorithm is one of the most popular student models in Intelligent Tutoring Systems (ITS) literature (Baker, Corbett and Aleven 2008). It is a Hidden Markov Model with two hidden states, 'learned' and 'unlearned,' representing the internal state of the student's mastery or lack thereof of a specific skill. It also has two observable states, 'valid' and 'invalid,' representing the validity of answers given by the student. $P(G)$ is the probability of a guess, $P(T)$ the probability of a skill being learned, and $P(S)$ the probability of a "slip," or a misuse of a known skill. $P(L_n)$ represents the initial likelihood a student knows skill n.

an individual student's needs, and (4) a graphical user interface, which may or may not include an on-screen agent character. See Figure 1.1 for an example ITS architecture consisting of these components; see Figure 1.2 for an example of a student model expressed as a Bayesian network. A broad overview of a variety of ITS system architectures can be found in a literature review by Nwana (1990). We describe several distinctions in ITS literature below that influenced our work in making robot tutors.

1.3.1 Model-Tracing vs. Curriculum-Sequencing Tutors

The first distinction that is important in our work corresponds to the two major families of automated tutors outlined by Desmarais and Baker (2012): (1) those that perform step-by-step guidance through individual problems in a given domain, called model-tracing tutors, and (2) those that perform curriculum sequencing to maximize a student's learning potential by choosing a path through the curriculum space, called curriculum-sequencing tutors. These two families have differing origins in the education literature, though they are not mutually exclusive in practice. The choice between them typically reflects the granularity of the student model of the tutoring system, whether the tutor is modeling a student's progress with specific steps to solve a certain category of problems, or the tutor is modeling a student's knowledge as he or she progresses through a problem space by picking the most appropriate problems to solve next.

In our robot tutoring work, we explore both approaches to automated tutoring. We created a model-tracing robot tutor for Chapter 5, which traces students' ability to perform steps in a cognitively-demanding puzzle solving task, where all of the puzzles are fixed in advance. For Chapter 6, we created a curriculum-sequencing robot tutor which chooses the most appropriate language-learning task for students among available tasks, based on an estimate of each student's skills related to those tasks. We find that the granularity of our modeling, whether within-problem as in model-tracing or between-problems as in curriculum-sequencing, reflected the intended length of time for the tutoring interactions we designed, such that model-tracing was more appropriate for



A spherical ball with a mass of 2.00 kg rests in the notch shown below. If there is no friction between the ball and the walls, what is the magnitude of the force exerted on the ball by wall1?

Answer:

T: There is a force acting on the ball at T0 that you have not yet drawn.
Explain further OK

T: Notice that the ball is supported by a surface: wall1.
Explain further OK

T: When an object is supported by a surface, the surface exerts a normal force on it. The normal

1. Fg_y+F1_y=0

How does the operating system interact with the word processing program when you create a document?

Student: the operating system allows you to save new information in a document

Tutor: I see uh
Tutor: Can you elaborate a bit on that?

Student: yes, the operating system creates space to save the document so that it is not lost when you open another program

yes, the operating system creates space to save the document so that it is not lost when you open another program

(A) User interface of ANDES, a workbook-style physics tutor.

(B) User interface of AutoTutor, a dialogue-driven computer science tutor.

FIGURE 1.3: Side-by-side comparison of the graphical user interfaces of Andes, a workbook-style Intelligent Tutoring System (ITS) (Schulze et al. 2000), and AutoTutor, a dialogue-driven natural language generating ITS (Graesser et al. 2008). ITS's that have animated or virtual agents produce better student engagement and satisfaction over workbook-style systems and may lead to better student outcomes (Lester et al. 1997; Prendinger et al. 2003).

shorter-term interactions and curriculum-sequencing was more appropriate for longer-term interactions.

1.3.2 Workbook-Based vs. Dialogue-Driven Tutors

Another important distinction in ITS literature that informs our work is the choice of the user interface for automated tutors. There are two dominant graphical user interface styles: what we call "workbook-style" tutors and "dialogue-driven" tutors. "Workbook-style" refers to a tutor interface that asks students to fill in the blanks as they work through a problem with the tutor. Typically, such tutors require students to show their work in great detail so that the tutor can better diagnose what each student knows and does not know. See an example of such an interface in **Figure 1.3a**.

The other dominant style is "dialogue-driven tutors," "character-driven tutors," or "conversational agents." In these ITS's, a student is expected to answer natural language prompts by typing in natural language statements to the tutor software. While solving problems, students are expected to produce a close equivalent of a series of teacher-written statements that define the key inferences or steps needed to solve a problem (**Rus et al. 2013**). The most significant such tutor is AutoTutor, see **Figure 1.3b** for a screenshot of its interface. AutoTutor does natural language processing to assess to what degree the content of a student's answer matches the key inferences needed to complete each problem (**Graesser et al. 2008**).

The distinction between these two popular interfaces, one typically with an on-screen character (i.e. dialogue-driven tutors), the other without (i.e. workbook-style tutors), allows us to ask how the presence of a virtual agent influences students in automated tutoring.

In the ITS community, the phenomenon of a student behaving differently in the presence of an on-screen character as part of the tutoring software is called the "persona effect" and its validity is debated (**Lester et al. 1997**). Most groups studying the persona effect find an increase in student attention, satisfaction, or motivation attributed to the presence of an on-screen agent (**Moundridou and Virvou 2002**; **Van Mulken, André and Müller 1998**), but only a handful of groups have found learning gain improvements as a result of these effects (**Baylor and Ebbers 2003**; **Prendinger et al. 2003**). This may indicate that the persona effect, or embodiment in robotics, only contributes to learning gains in some domains but not others. Conversational agents are becoming more popular in the ITS community according to a recent survey of such agents by **Rus et al. (2013)**, so soon we may know more about the persona effect and whether we can effectively harness it in real-world tutoring applications.

In **Chapter 2** we present results on the benefit of having a physically-embodied robot tutor compared to an on-screen virtual agent and a disembodied voice. We find that physical presence leads students to pay more attention and improve their learning outcomes relative to the other two conditions.

1.3.3 Personalization in Automated Tutors

The personalization a tutor does to match the needs of each student is what accounts for the relative success of one-on-one tutoring over group instruction in traditional classroom settings (**Merrill et al. 1992**). Personalization is a feature of all automated tutoring systems and many kinds of personalization have been pursued by ITS researchers – from inferring a student's motivation based on his or her facial expressions or posture (**Conati and Maclaren 2009**; **D'Mello 2012**), to detecting if students are trying to abuse the hint and help features of ITS's to game the system to improve their scores (**Baker, Corbett and Koedinger 2004**).

The most significant type of personalization in automated tutors is the student model (**Hogan and Pressley 1997**). An example Bayesian Network student model is found in **Figure 1.2**. We offer several preliminary student models for robot tutors in **Chapter 5**

FIGURE 1.4: RUBI is a humanoid robot designed to interact with young children, 18 to 24 months old. It has articulating arms, an expressive face, and a touch-screen tablet-like midsection on which it displays educational content. The RUBI platform is best known for a study of its use to teach vocabulary in a preschool classroom setting (Movellan et al. 2009).

and Chapter 6, targeted towards a variety of learning tasks and student populations. In addition to this form of personalization, we also explore the role of affect in human-robot tutoring in Chapter 4.

1.4 Robot Tutoring

Perhaps the most developed robot tutoring platform is from a project called RUBI. The RUBI project began in 2004 and is now in its fifth hardware iteration; for an overview of the project see Movellan et al. (2007). RUBI is a humanoid robot designed to interact with 18- to 24-month-old children. It has articulating arms, an expressive face, and a touch-screen tablet-like midsection on which it displays educational content. See **Figure** 1.4 for a picture of the robot interacting with children. The RUBI platform has been used for a variety of studies, from teaching vocabulary words in a preschool classroom setting (**Movellan et al. 2009**) to detecting children's preferences for different activities in a simulated home setting (**Malmir et al. 2013**). **Ruvolo et al.** (**2008**) used RUBI to perform apprenticeship learning in which the robot learned to teach from demonstration by a human teacher. The studies with RUBI have not focussed on the personalization aspect of tutoring, which is what we look at in this dissertation.

The other dominant family of robot tutoring is those that act as telepresence agents for teachers who operate the robot from a distance and conduct either one-on-one tutoring or traditional classroom group instruction. (**Hyun, Yoon and Son 2010**) established some experimentally-derived guidelines for using robots remotely in classrooms, done with the Korean robot tutor called iRobiQ, a humanoid similar in design to RUBI. Another Korean project, Engkey, is specifically designed for English-language tutors who may live abroad (**Yun et al. 2011**). To see Engkey in use in a classroom setting, see **Figure** 1.5. Telepresence robots for education are a burgeoning technology, with other notable examples being MIT's Huggable robot (**Lee et al. 2008**) and another Korean project called ROBOSEM (**Park et al. 2011**). All of these projects, however, assume a human teleoperator who is responsible for the instructional content. In our work, although we do occasionally use human operators for some aspects of an interaction, the educational content is always automated and independent of the human operators.

FIGURE 1.5: 'Engkey' is a telepresence robot agent for teachers and tutors to operate from a distance (Yun et al. 2011). It uses a video feed of the human tutor's face to convey affect information. This family of robot tutors, unlike our work, requires a human teacher to provide the instructional content.

1.4.1 Personalization in Human-Robot Interaction

Though we are the first group to look at personalization in robot tutoring, we are not the first to look at personalization across all of robotics. The most significant previous work in personalization is Snackbot, a robot that personalized dialogue in reference to an individual user's history of snack choices (i.e. an apple or a candy bar), was found to be more engaging by users than a non-personalized version of the same robot, leading to an increased desire to use the robot and an increase in social behavior directed toward the robot (Lee et al. 2012). A robot weight loss coach by Kidd and Breazeal (2008) generates customized dialogue based on the progress of the user but their research does not isolate the effect of personalization. In other work, a long-term study of elementary students playing chess with a robot explored how supportive the students perceive the robot tutor to be depending on the kind of feedback it gave students (Leite et al. 2012). In the work of Sung, Grinter and Christensen (2009), users that decorated, and thus "personalized," their Roombas, self-reported higher engagement with the robot and more willingness to use the robot in the future.

Previous work in personalization in robotics research is varied, but we look specifically at how personalization affects robot tutoring interactions. We find out to what extent personalization influences students' perception of the robot and, more importantly, to what extent personalization impacts the learning gains made by students.

1.5 Dissertation Overview

This dissertation answers a foundational question in robot tutoring, delivers experimentally-derived design guidelines for future work in robot tutoring, and provides novel robot tutoring personalization systems that improve student learning outcomes over non-personalized systems by 1.2 to 2.0 standard deviations, corresponding to gains in the 88th to 98th percentile.

1.5.1 Foundational Question: "Why Use a Robot?"

We first answer a foundational question in the field of robot tutoring: "Why use a robot?" Our work shows that the physical presence of a robot can improve student learning outcomes over on-screen character tutors by as much as 0.3 standard deviations, corresponding to gains in the 62nd percentile.

The presence of on-screen characters in automated tutoring systems have been shown to improve student engagement, satisfaction, and, in some studies, student learning outcomes over automated tutoring systems that do not have on-screen characters (**Baylor and Ebbers 2003**; **Lester et al. 1997**; **Prendinger et al. 2003**). We investigate whether the physical presence of a robot tutor can have a similar or perhaps stronger effect than the presence of on-screen characters in automated tutoring. We compared three conditions in our investigation, each of which received the same instructional content: one in which the content was delivered by a robot tutor, one in which the content was delivered by an on-screen character tutor based on the robot in the first condition, and one in which the content was delivered by the same voice as in the first two conditions, but with no physical or virtual embodiment. We found that the physical embodiment of the robot increased students' attention and students who received robot tutoring learned significantly more of the instructional content than those in the other two groups. We conclude that the physical embodiment of a robot can be leveraged to improve student learning outcomes over on-screen character tutors. This work is described in **Chapter 2**.

1.5.2 Experimentally-Derived Design Guidelines

To maximize the impact of our automated tutoring systems, we first assessed the key features of expert human tutoring behavior. Understanding what makes an expert human tutor effective is not as straightforward as it may seem. When education researchers study human-human tutoring, a major potential confounding variable in their work is the "chemistry" between tutor and student, which determines how effectively they are able to communicate (Topping and Ehly 1998). This effect limits researchers' ability to compare one tutor's behavior to another, even when they teach the same student, and as a result it is difficult to generalize about the nuances of successful tutoring practices.

Our work overcomes this limitation by using robots as students paired with human tutors. Unlike human students, robot students can produce the exact same behavior in multiple instances and with different human tutors. Having consistent robot reactions to a variety of human tutoring behavior allows us to investigate the commonalities and differences between the human tutors more precisely. Based on these investigations, we provide design principles for future work in automated personalization systems for robot tutoring.

- Does the relative successfulness of a student influence the kinds of tutoring a human tutor provides? We found that when the same human tutor teaches two robot students, one a more successful student and the other a less successful student, humans provide significantly different feedback to these types of students. When teaching a less successful student, human tutors give feedback much earlier in each task and more frequently throughout the task. Human tutors also vary the affective content of their instruction to these robot students. When teaching the robot student that makes more frequent mistakes, human tutors provide significantly more affect in their instruction, the majority of which is encouraging and motivational. Whereas, when they teach a more successful robot student, we found that human tutors provide less and less feedback over time. These findings highlight the importance of treating more-successful and less-successful students significantly differently in automated tutoring, something that is not currently done in many automated tutoring systems. This work, along with resulting design guidelines, can be found in **Chapter 3**.

- In the work described above we found that human tutors provide very different kinds of instruction to students that differ in their performance on learning tasks, but would human tutors personalize their instruction between students who perform identically on learning tasks? We investigate this in a study with three conditions, all of which perform the learning tasks identically and receive identical scores, but each of which have distinct patterns of emotional responses. Either the robot responds to the scores it receives with: (1) emotionally-appropriate responses such as, "That was great!" for good scores or, "I am so sad," for poor scores, or (2) often emotionally-inappropriate responses, such as "We did amazing!" for poor scores, or (3) apathetic responses such as, "That was OK." We found that human tutors do personalize their instruction to students of exactly the same learning task performance, based on the students' responses alone. We found that the robot students who gave feedback that was often emotionally-inappropriate or apathetic caused human tutors to be disengaged with the teaching process, evidenced by their performance of fewer demonstrations with less enthusiasm and accuracy than participants in the emotionally-appropriate group. We conclude from these findings that human tutoring personalization goes well beyond a learner's task performance and that the affective content of a tutoring dialogue is of critical importance to human tutors. This work, and resulting design guidelines, can be found in **Chapter** 4.

1.5.3 Robot Tutoring Personalization Systems

This dissertation contributes two kinds of novel systems for robot tutor personalization, one intended for shorter-term robot tutoring interactions and one for longer-term interactions.

- We created a model-tracing robot tutor that teaches adults to play a cognitively challenging puzzle game called 'Nonograms' or 'Nonogram puzzles.' While participants solve a series of Nonogram puzzles, the robot tutor assesses their skill competency on a 10-skill Nonograms puzzle-solving curriculum we authored. The robot gives step-by-step advice several times during an interaction, much like it would if it were tutoring in math or physics, based on one of an individual student's weakest skills. Participants who received personalized lessons from the robot tutor improved their puzzle-solving time an average of 1.2 standard deviations over participants who received non-personalized lessons, corresponding to gains in the 90th percentile. This result validates the effectiveness of our personalization system, confirming that the lessons we chose for each student were significantly better suited to them than those in the non-personalized condition. A description of this work can be found in **Chapter** 5.

- We also created a longer-term personalization system based on a Hidden Markov Model (HMM) that learned its transition probabilities over the course of a two-week-long tutoring interaction, teaching an English as a Second Language (ESL) curriculum to native Spanish-speaking first graders. In this work we implemented a curriculum sequencing tutor to maximize a student's exposure to unfamiliar or forgotten English grammar skills. The students who received personalized instruction outperformed students who received non-personalized instruction in a post-test by an average of 2.0 standard deviations, corresponding to gains in the 98th percentile. This work can be found in **Chapter 6**.

We conclude this dissertation with a summary of our contributions in **Chapter 7**.

Chapter 2

"Why a Robot?": The Role of Embodiment in Robot Tutoring

In this chapter we address a fundamental question in robot tutoring: "Why use a robot?" We show that the physical presence of a robot tutor has an effect on students that can be leveraged to increase learning gains by 0.3 standard deviations over on-screen character tutors, corresponding to learning gains in the 62nd percentile.

In order to investigate the effects of embodiment in automated tutoring, we designed an experiment consisting of three tutoring conditions with differing embodiments, holding the instructional content constant between the three conditions. Participants either received lessons from: (1) a robot tutor, (2) an on-screen character tutor, based on video footage of the robot in the first condition, or (3) a voice-only tutor with no physical or virtual embodiment, which used the same voice as in the previous two conditions.

The domain we chose for this learning task was a cognitively challenging and relatively obscure puzzle game called 'Nonograms', in which players make progress in each puzzle by making logical inferences about a set of constraint satisfaction problems. Choosing this relatively complex pedagogical domain allows us to better isolate the effect of the embodiment on student outcomes, rather than choosing a simpler pedagogical domain, like a vocabulary memorization task, where simply engaging with a robot may increase students' willingness to practice and thereby lead to learning gains. We found that participants who received tutoring from the robot learned to solve Nonograms better than in the other two groups and improved their same-puzzle solving time significantly over participants in the other groups. We conclude that the effects of physical embodiment can produce student learning gains in an automated tutoring interaction, even for adults engaged in complex learning tasks.

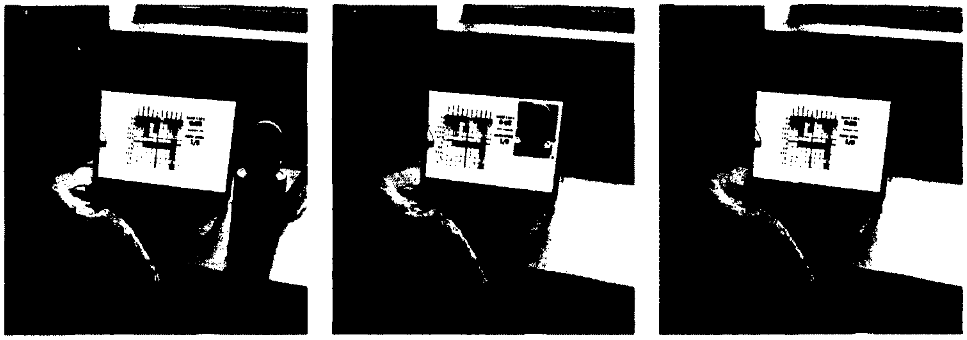
2.1 Related Work

Though we are the first to investigate the effect of physical presence on the success of automated tutoring systems, researchers in the Intelligent Tutoring Systems (ITS) community have investigated the effect of on-screen characters on the success of automated tutoring systems. The phenomena of students behaving differently in the presence of an on-screen character is known as the 'persona effect' in ITS literature (Lester et al. 1997). Research on the persona effect has found that the presence of an on-screen character increases student attention, satisfaction, or motivation over similar agentless automated tutoring systems (Moundridou and Virvou 2002; Van Mulken, André and Müller 1998).

However, only a handful of studies have found that these increases in student attention, satisfaction, or motivation led to increased learning gains (Baylor and Ebbers 2003; Prendinger et al. 2003). These results indicate that the persona effect influences students but that the presence of an on-screen character does not, in and of itself, guarantee improved learning gains. We discuss this further in Chapter 1, Section 1.3.2.

Perhaps the physical presence of a robot tutor can engender more trust, compliance, motivation, or engagement than the two-dimensional presence of an on-screen tutor. If so, we may be able to use those effects to improve student learning outcomes. The effect of the physical presence of a robot in human-robot interactions has been investigated in teamwork, therapy, and coaching domains, though not yet in automated tutoring. There are two types of results in this work: changes in self-report measures as a result of embodiment and changes in compliance as a result of embodiment. We summarize these below.

- A significant result among the self-report measures was found by Kidd and Breazeal (2004), in which a physically embodied robot was rated by participants as more enjoyable, more credible, and more informative than an on-screen character in a block-moving task. In Wainer et al. (2007), an embodied robot was rated by participants as more attentive and more helpful than both a video representation of the robot and a simulated on-screen robot-like character. Tapus, Tapus and Matarić (2009) find that individuals suffering from cognitive impairment or Alzheimer's disease reported being more engaged with a robot treatment than a similar on-screen agent treatment.



(A) Experiment apparatus in the *robot tutor* condition.

(B) Experiment apparatus in the *on-screen tutor* condition.

(C) Experiment apparatus in the *voice-only tutor* condition.

FIGURE 2.1: Experiment apparatus by condition.

- Compliance results include Kiesler et al. (2008), in which participants who received health advice from a physically-present robot were more likely to choose a healthy snack than participants who received the same information in robot-video or on-screen agent conditions. Bainbridge et al. (2008) found a significant improvement in the compliance of participants to a robot's requests to throw away books in a physically-present robot versus a video representation of the same robot.

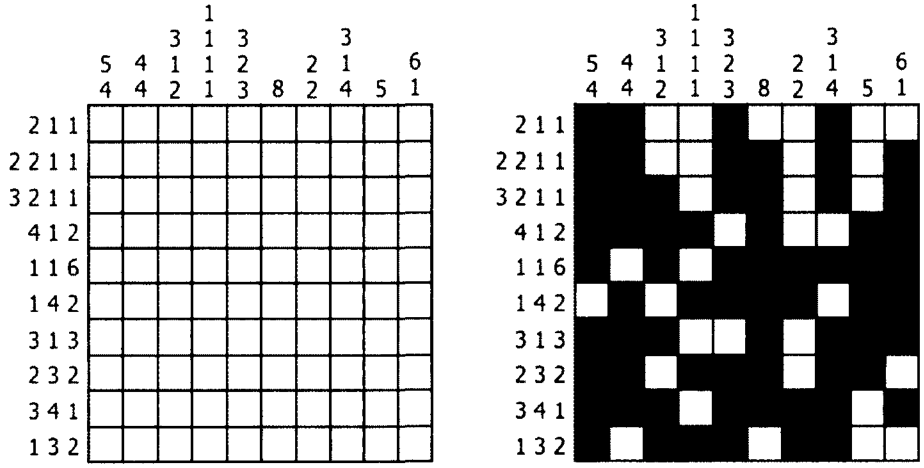
We use task-performance measures in our work, in the form of Nonograms puzzle-solving time, as well as self-report measures, in the form of exit surveys, to investigate the effect of the physical presence of a robot in an automated tutoring interaction.

2.2 Overview

In this experiment participants were asked to solve a series of four logic puzzles called "Nonograms." Periodically, as participants were solving these puzzles, the tutor interrupted them to demonstrate a puzzle-solving skill relevant to the specific puzzle they were solving. These puzzle-solving lessons consisted of pre-recorded audio with synchronized lesson-specific on-screen visual aids, each between 21 – 47 seconds long, and each describing a unique skill. These lessons were delivered to participants in one of three ways, depending on the experimental condition the participant was randomly assigned to: either by (1) a robot tutor, (2) an on-screen character tutor, or (3) a voice-only tutor with no physical or virtual embodiment. The apparatus for the each condition can be found in **Figure** 2.1. The faster a participant was able to solve the puzzles, the better at puzzle-solving we judged them to be. We compare the mean puzzle-solving times between participants across groups to evaluate the effect of an automated tutor's embodiment on student learning outcomes.

2.3 Curriculum: 'Nonograms'

To minimize the potentially biasing effect of differences in participants' prior knowledge, we chose a pedagogical domain which was likely to be unknown to participants. 'Nonograms,' also called 'Nonogram puzzles,' are a Japanese grid-based fill-in-the-blanks game similar to Sudoku. Nonogram puzzles are a difficult cognitive task, one that requires several layers of logical inferences to complete. Solving a Nonogram puzzle of arbitrary



(A) Sample Nonogram puzzle, blank.

(B) Sample Nonogram puzzle, solved.

FIGURE 2.2: A sample Nonograms puzzle. The objective of Nonograms is, starting with a blank board like in **Figure 2.2a**, to find a pattern of shaded boxes on the board such that the number of consecutively shaded boxes in each row and column appear as specified, in length and order, by the numbers that are printed to the left of each row and above each column like in **Figure 2.2b**. For a more detailed explanation see **Section 2.3**.

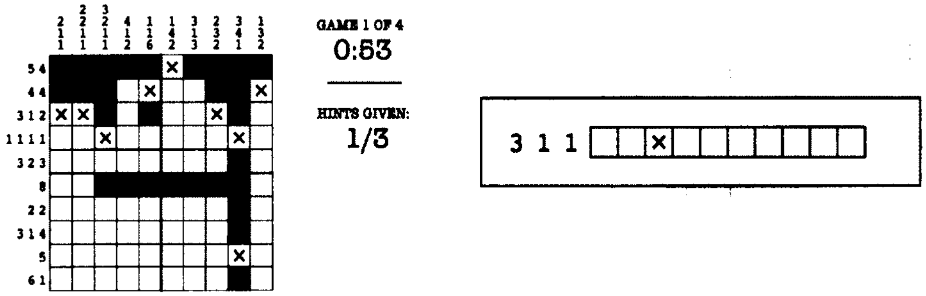
size is an NP-complete problem (**Nagao and Ueda 1996**), meaning that no efficient computational solution is known. An example of a Nonogram puzzle with its solution can be found in **Figure 2.2**.

The objective of Nonograms is, starting with a blank board, to shade in boxes on the board such that the number of consecutively shaded boxes in each row and column appear as specified, in length and order, by the numbers that are printed to the left of each row and above each column. For instance, a row marked as '4 2' must have 4 adjacent shaded boxes, followed by 2 adjacent shaded boxes—in that order, with no other boxes shaded in that row, with at least one empty box between the sets of adjacent shaded boxes, and with any number of empty boxes before or after the pattern. We refer to these contiguous sets of shaded boxes as 'stretches.' For instance, the row described above requires two 'stretches,' one of length 4, the other of length 2. One solves the puzzle when one finds a pattern of blank and shaded boxes such that all of the requirements for each row and column are satisfied. See **Figure 2.2a** and **Figure 2.2b** for a sample puzzle and its solution.

In a typical puzzle, one cannot solve most rows or columns independently. Instead, one must infer the contents of parts of rows or columns and use previous inferences as the basis of subsequent inferences. This is the case because when you shade in a single box on the board, you affect both its row and its column. That affects the rest of that row or column, and the rest of that row or column can affect any of the rows or columns that it intersects. One must make each move without violating any of the row or column requirements of intersecting rows and columns.

One way to make progress in Nonograms is to shade boxes that the player infers must be shaded, regardless of how the rest of the row or column is shaded. Another way is to infer that a box or a set of boxes cannot be shaded. When participants made such an inference they marked that box or those boxes with a red 'X' symbol. These 'X's can be seen in the screenshots of the graphical user interface **Figure 2.3** as well as in the examples provided in the lessons, documented in **Section 2.4** below.

We created a full-screen Nonograms computer program that participants used with a mouse and keyboard. The user interface provided a timer and a count of how many



(A) Screenshot of our Nonograms graphical user interface, during gameplay. The red 'X' marks are flags participants can set when have determined not to shade in a specific box. The robot tutor encourages use of these flags.

(B) Screenshot of our Nonograms graphical user interface, during a lesson (called "hints" during the experiment). The overlay here is visual aid that the robot tutor used to teach lessons. For example, in this case, the robot instructs the participant to put an 'X' in the first two boxes.

FIGURE 2.3: Screenshots of our Nonograms graphical user interface.

lessons (called "hints" in the interface) the participant had received and how many they would receive; screenshots of which are found in **Figure 2.3**.

Participants were asked to play four puzzles on ten-by-ten grids with a time limit of fifteen minutes per puzzle. Every participant in this study was asked to solve the same sequence of four puzzles. The first puzzle was the easiest, though not a trivial one. The second was incrementally harder and the third was harder still.

The fourth puzzle, however, consisted of the same board as in the first puzzle, but disguised by a rotation of 90°, such that the column stretch requirements were swapped with row stretch requirements and vice versa. This meant that the first puzzle and the last puzzle were of the same difficulty and required the same knowledge to complete. This manipulation enables us to make within-subjects comparisons about the extent to which each participant improved the skills necessary to solve the first puzzle over the course of their participation in this study.

Figure 2.2, the sample Nonogram puzzle above, is the same puzzle participants solved first. In the fourth puzzle, the rows and columns were swapped.

2.4 Skills & Lessons

In this study, the tutor interrupted the participant three times per puzzle. The puzzle game was paused and the tutor delivered a short Nonograms lesson. The lessons ranged from 21 seconds to 47 seconds in length and consisted of a voice recording and a set of animations presented on screen, overlaid atop the paused Nonograms puzzle interface.

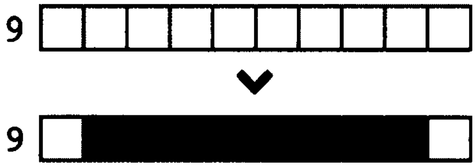
Ten Nonograms puzzle-solving skills were identified based on the author's subjective experience of Nonograms. These ten skills are not universally identified rules or strategies for Nonograms, but rather a set of mutually exclusive row or column patterns in which one can logically fill in some of the remaining empty boxes.

For example, a stretch of length 9 can fit in a blank row or column of 10 boxes in only two ways. Either it fills the first box and 8 more, or it fills those same middle 8 boxes and the last box. In either case, the middle 8 boxes are shaded. We generalize this phenomena as one of ten skills we identified, such that for an empty row or column with just one stretch requirement of n where $n > 5$, the middle $(2n - 10)$ boxes are shaded. This is one example skill, corresponding to the first of the ten lessons. The text of all ten lessons, as well as the visual aids offered to participants during the lessons can be found below.

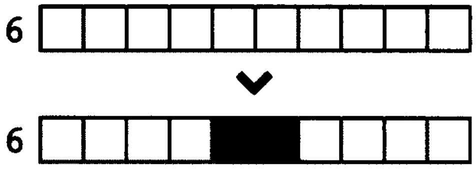
The following is the exact text of the Nonograms puzzle-solving lessons spoken by the tutor, along with the visual aids provided during each lesson.

2.4.0.1 "Shade Middle Boxes of Long Stretch"

"If a stretch has just one big number, I bet some squares will be shaded no matter where we put that stretch. Take a look at this example. The only stretch in this row is 9, but no matter where you put it the middle 8 squares will be shaded."

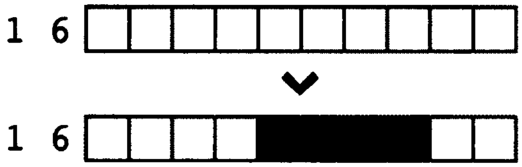


"I think this works for any row with one long stretch like here, where the only stretch is a 6. We can put it all the way on the left, all the way on the right or anywhere in between. But the middle squares get shaded either way, so we can confidently mark them down without guessing."

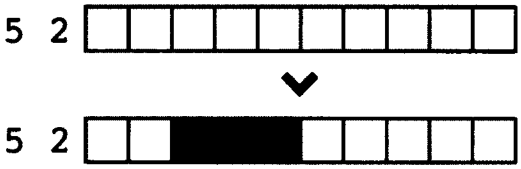


2.4.0.2 "Shade Middle Boxes of Long Stretch, Advanced"

"Here's another hint for you. Take a look at this row. There's only a one and a six and the one has to come first. That six is a long stretch and there are only a few options for where to put it. Imagine that the one is all the way at the beginning of the row. If the one is there then we can figure out how to shade some of the other boxes that the six will have in common, no matter where it is. The six can be as close as possible to the one, or as far away as possible. Here are the four squares those two cases have in common."



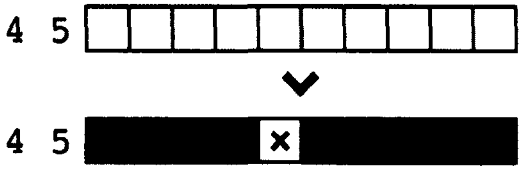
"In this other example, now the longer stretch comes first. Imagine that the '2' is at the very end of the row, then the '5' can be as far away from the '2' as possible, as close as possible, or anywhere in between, yet some squares will always be shaded."



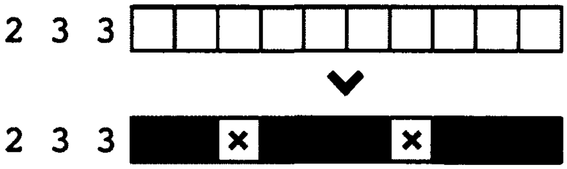
2.4.0.3 "Completing a Row"

"I've noticed something that should help you get started on empty rows. If there's only one way to fill in a row then that's the way it has to be. It sounds obvious, but take a look at this example. Starting from the left, we shade four blocks, put an 'X', then shade

five more blocks. There's no way we can arrange them that fits the number pattern on the side so we can mark it down confidently."

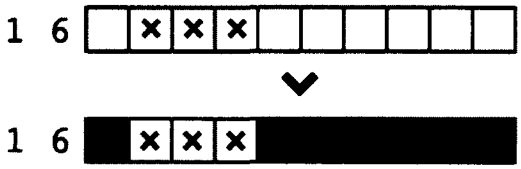


"It even works for rows with more than two stretches. See here, we start from the left then shade two, then an 'X', shade three, then an 'X', and then shade three more. The row begins and ends with the first and last stretches, and there's only room for one 'X' in between the stretches, which means there aren't any other possible configurations."

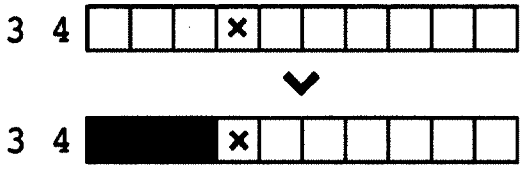


2.4.0.4 "Completing a Row, Advanced"

"I have another tip that I think might help. If you start off with a few boxes in any row that are marked with 'X's or already shaded, you may be able to fill out the whole rest of the row based on what requirements are left. For instance, in this example, the three blocks that are X'ed out means that the '1' and '6' stretches can only go in one way."

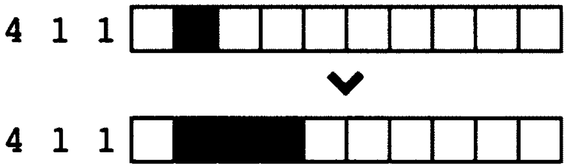


"This also works for partial rows or columns. As you can see in this row, there are two open areas, so you know that the '3' has to be one one side and the '4' needs to be on the other side of the X's. You can shade the three empty boxes confidently."

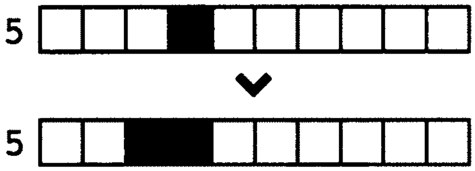


2.4.0.5 "Shade to the Lth box."

"This strategy is kind of tricky, but I think it can help you. Look at the first stretch of a row and call it's length 'L'. If a box is shaded in the first L boxes of a row, you know that that box has to be part of that first stretch, so you can shade up to the Lth box. Up the Lth box will be shaded in no matter where the first stretch is placed. Take this row for example. The first stretch is a '4', and the second square is shaded, so we know that it has to be part of the first stretch. Therefore we can shade up to the fourth square."

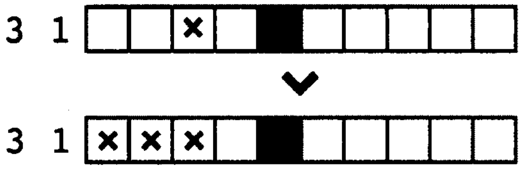


"In this next example, the first stretch is a five and the fourth square is shaded. So, we fill up all the squares from four to five, which in this case is just the fifth square."

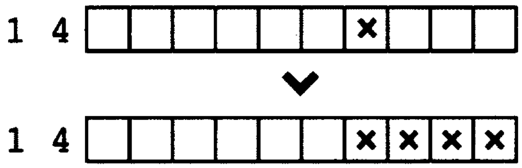


2.4.0.6 "X-Out Beginning/End If It Will Not Fit First/Last Stretch"

"Here is a strategy to help you X-out more boxes. Take a look at this example. The first stretch is a three, but the third box is X'ed out, so there isn't enough room to put it at the beginning. That means we can X-out all the other boxes leading up to it, since the first stretch in this row must be three boxes long."

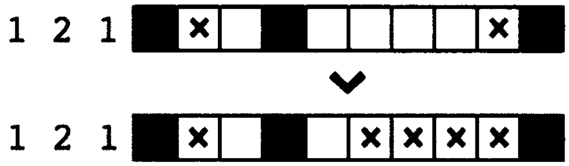


"We can also do this at the end of a row too. The last stretch is a four, but the third square from the last is X-ed so we know that all the other squares after it have to be X-ed out too."



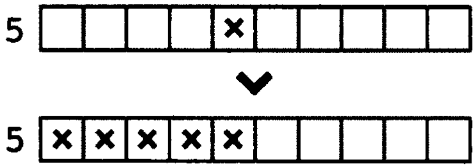
2.4.0.7 "X-Out Boxes Out Of Reach"

"Here is a handy way to finish rows or columns that you've already made progress on. If there is only one stretch left in a row, and you know where it is, you can X-out any boxes that are too far away to be part of the stretch. Take this row for example. The only unfinished stretch is of length two and you know that it has to be around the fourth square, so you can X-out blank spaces farther away, like the blank sixth, seventh, and eighth squares."

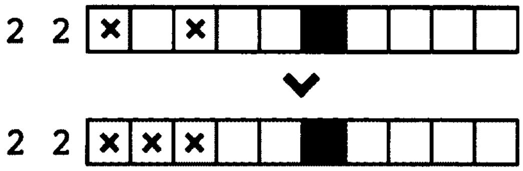


2.4.0.8 "X-Out Boxes That Don't Fit Smallest Stretch"

"If you see any series of blank squares bordered by X's and those squares are shorter than the shortest stretch, you can go ahead and X them all out. This is because stretches are always bounded by X's or the end of the board, and if there isn't enough space for even the smallest stretch then there can be no shaded squares there. In this case the smallest and only stretch is a five but we see a series of bounded boxes of length four, so we know the five cannot fit."

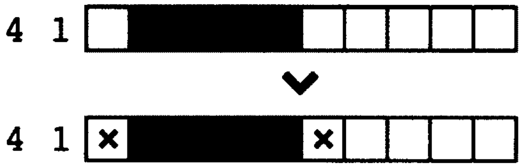


"Now, in this other case, the smallest stretch is a two, so if we see a stretch of one blank boxes bounded by X's we can X it out."



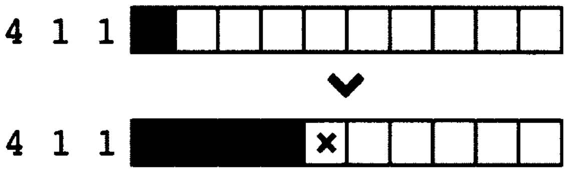
2.4.0.9 "X-Out Boxes Around Max Length"

"Here's something that might be able to help. If in any row or column you see a bunch of shaded boxes in a row and the number of boxes is equal to the longest stretch, you can put X's around them. Let's assume that along the way we've shaded these four boxes. Since the longest stretch in the row is four, we know that there have to be X's at the ends of these boxes."

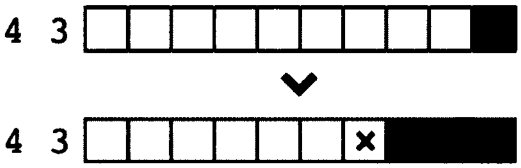


2.4.0.10 "If First or Last Box Shaded"

"I thought of a good technique you could use. If the first square of a row or column is shaded, then it must be part of the first stretch. You can fill in the rest of that stretch and put an X on the end. See, here we know that first square is part of the '4' stretch, so we can complete the stretch and put an 'X' at the end."



"This strategy also works for the last square of a row or column too. If the last square is shaded, it's part of the last stretch, so you can work backwards to fill in more squares like this one. The last stretch is a '3', so we can shade the last three boxes then put an 'X'."



Participants received three lessons per puzzle, for each of four puzzles. On average, 3.2 of these lessons were repeated per participant, depending on the lesson ordering. Lessons were given either when a participant made no moves for 45 seconds or as he or she filled the 25^{th}, 50^{th} or 75^{th} box on the board (of 100). The user interface displayed the number of lessons remaining for each puzzle at all times.

The lesson ordering was based on a personalization algorithm we designed. The details of this algorithm and an evaluation of its effectiveness are the subject of **Chapter 5**. The personalization ensured that the lessons the robot gave corresponded to the skills in which each participant had the least competency, among those that were applicable to the current game state. In all three conditions in the study, the personalization system was held constant. If two participants had made the exact same moves, no matter what experimental conditions they were in, they would have received the exact same lesson ordering.

2.5 Conditions

In all three experimental conditions in this study, the tutor spoke in the same voice and used the same pre-recorded lessons. The tutor served three roles.

- First, it refereed the puzzle game. It welcomed participants when they started, told them when they had finished or when they had run out of time, and told them when the experiment was over.

- Second, to discourage the participants from explicitly asking the tutor for advice, in the physical robot tutoring and on-screen tutoring conditions, the tutor turned its head to follow the mouse cursor when the participant was solving a puzzle. The tutor did not respond to any direct queries.

- Third, the tutor delivered three lessons per puzzle. The tutor started each lesson by saying "I have an idea that might help you," or "Here is another hint for you," and in the case of the physical and virtual tutors, it would turn from the screen to face the participant and bounce up and down to indicate it was talking to the participant. In the voice-only condition, the tutor simply spoke the lessons with no physical or virtual representation. When the tutor repeated a lesson, it would apologize first (i.e., "I'm sorry to repeat this hint but I think it might help.").

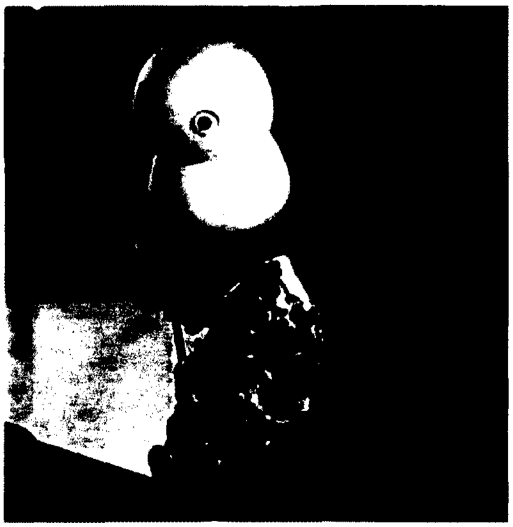


FIGURE 2.4: Keepon is an 11-inch tall, stationary, yellow, snowman-shaped robot with small, round eyes, one of which contains a camera, and a small, round nose containing a microphone. Its motors, though visible here, are typically enclosed in a black metal cylindrical tube. Keepon can rotate left and right, lean side to side, tilt up and down, and bounce up and down.

2.5.1 Robot Tutor

For the robot tutor condition, we used a robot called Keepon. Keepon is an 11-inch tall, stationary, yellow, snowman-shaped robot with small, round eyes, one of which contains a camera, and a small, round nose containing a microphone. The robot can be seen in **Figure** 2.4. Keepon stands atop a 12-inch black cylindrical pedestal which houses its motors and controllers. It has four motors that allow it to rotate left and right, lean side to side, tilt up and down, and bounce up and down. Keepon was originally developed for use with children with social skill deficiencies (**Kozima, Nakagawa and Yasuda 2005**).

Instead of using a vision system to detect the state of the gameboard by processing input from the robot's camera, we simply allowed the tutor in all three conditions to query the graphical user interface program directly to receive perfect knowledge of which boxes were shaded and which were not. This was done to ensure that the tutor's knowledge of the state of the gameboard was consistent across all three tutoring conditions.

2.5.2 On-Screen Tutor

The on-screen character tutor was comprised of video recordings of the physical robot tutor, played to match the behavior of the physical robot in the physical robot tutoring condition. We recorded four videos: (1) the robot facing the computer, watching the gameboard as a participant solves a puzzle, (2) the robot turning around to face the participant, (3) the robot "talking" to the participant by bouncing up and down, and (4) the robot turning back around to face the gameboard on the computer. We recorded these videos in such a way that they would appear seamless when played in this order. We thereby created an on-screen representation of what participants experienced in the physically-embodied robot condition.

2.5.3 Voice-Only Tutor

The voice-only tutor used all of the same voice recordings as in the other two conditions, but with no physical or virtual representation.

2.6 Participants

There were 60 participants in this study, between 18 and 35 years of age, most of whom were undergraduate and graduate students of Yale University. There were 20 participants in each group of the three groups. Exclusion criteria for participants were lack of English fluency, prior academic experience with robotics or artificial intelligence, and prior experience with Nonograms, the most common form of which in the U.S. is a Nintendo game called, 'Picross DS' (Nintendo 2007*a*).

2.7 Procedure

Participants watched a five minute instructional video about the graphical user interface and read a two-page instruction manual we authored that describes the rules of Nonograms. In the instructional content, we encouraged participants to use logical reasoning to make moves rather than guessing. Excessive guessing can disrupt the personalization of the tutor and can confuse novice Nonograms players potentially causing a loss of data. After watching the video and reading the instructions, participants were given an opportunity to ask clarifying questions about the rules and the interface. The text of the instructions given to participants appears in full in Section 2.8 below.

During the experiment, participants were alone in a room with the computer, and the robot in the robot tutoring condition, as well as a video camera positioned behind them; see Figure 2.1. Participants chose when they were ready to start each new puzzle. Each round ended either when the participant solved the puzzle or when the fifteen minute time limit had elapsed.

After the conclusion of the final puzzle, participants in the three groups that received tutoring were asked to complete a survey consisting of five Likert-scale questions designed to assess whether the tutor's lessons were helpful, clear, and influential, as well as the participants' perceptions of the tutor. We asked participants to respond to the following prompts: "How relevant were the tutor's lessons?", "How much did the tutor's lessons affect your gameplay?", "How well did you understand the tutor's lessons?", " How smart/intelligent do you think the tutor is?", and " How distracting/annoying did you find the tutor to be?" Participants responded with a score from 1 to 7, 1 being "not at all" and 7 corresponding to "very much." The intention of these questions was to reveal differences between tutoring conditions that could explain any performance differences between groups.

2.8 Instructions

The following written instructions, along with sample Nonogram puzzle and solution, were given to participants before beginning their participation in this study.

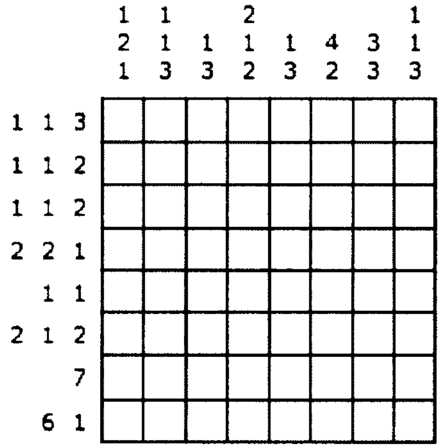
This is a study about the effectiveness of digital tutors.

You will play a puzzle game called "Nonograms" and an on-screen tutor will give you advice to help you get better as you play. Several times per game, the tutor will teach you *a puzzle-solving strategy that you may not have tried yet. The following are the rules of Nonograms.*

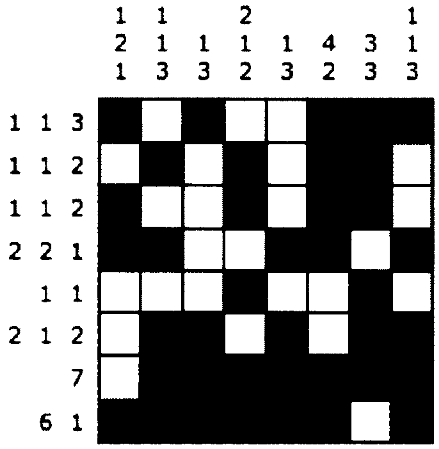
Nonograms is played on a ten-by-ten grid. Your task is to figure out which of the boxes on the grid need to be shaded and which don't. You'll see a series of numbers to the left of every row and at the top of every column. These numbers are different for every puzzle and provide constraints for which boxes can be shaded and should not. The numbers are called "stretches"; they tell you how many consecutive boxes are shaded in each row and column. For example, if the stretches for a row are "2 3" then 5 total boxes are shaded in that row. But not just any 5 of the boxes. First there must be two contiguous boxes shaded somewhere and after that three consecutive boxes are shaded. The stretches are listed in the order that they must appear in that row or column. A puzzle is solved when all rows and all columns have all of the stretches they are required to have and no extra boxes are shaded.

All Nonogram puzzles start out entirely blank. To play, mouse over the box you want to shade in, and press the space bar to shade it. Also, you can put an X in a box that you think won't be shaded in. To do that, press the X button. If you want to clear a box and make it blank again – mouse over it and press 'C'.

An important thing to know about Nonograms is that you will never need to guess in order to make progress in the game. As you look around the board, you will always be able to find some part of the board that you can fill in. Sometimes you'll be able to fill in an entire row or column at once, but more often than that you'll have to use the process of elimination to determine which boxes in a row must be shaded and which definitely



Sample puzzle, blank.



Sample puzzle, solved.

aren't shaded. To do that, we encourage you to imagine all the possible configurations of a row's or column's stretches and decide whether any boxes are always shaded or always X-ed out, in every possible configuration. Once you've started to make progress on the board, you will be able to use the boxes you've already filled in to help you fill in more boxes. If you fill in several boxes in a row, you can check the columns those boxes were in to see if that new information helps you determine something about the boxes in those columns.

By deducing each move, logically, from the board and from previous moves, you'll be able to go from a blank board to a completed board in no time. The more experience you have with this game, the better you will do.

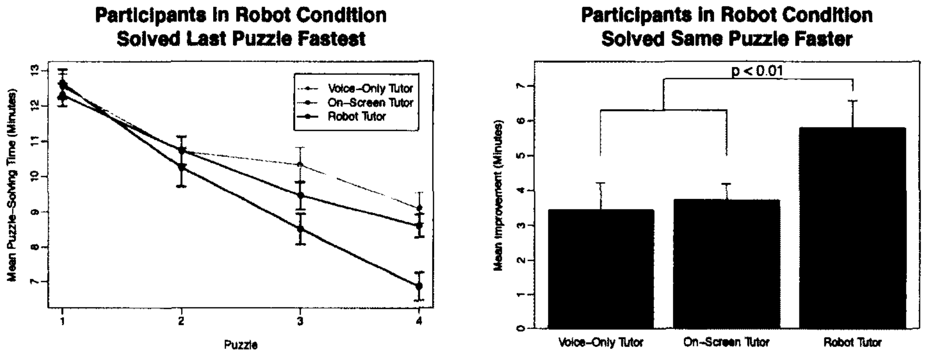
Those are all the rules for Nonograms. We ask you to play 4 boards, back-to-back. We want you to finish them as quickly as you can. If you're not done in 15 minutes, the program will move you on to the next puzzle. We encourage you to work as hard as *you can on each puzzle, and please remember that you shouldn't need to guess to make progress.*

Good luck!

2.9 Results

This study investigates the effect of embodiment on learning gains in automated tutoring systems. We measure the length of time participants needed to solve each of the four Nonogram puzzles. Lower puzzle-solving times are considered better puzzle-solving performance and indicate better Nonograms puzzle-solving skill competency. If a participant did not complete the puzzle in the allotted fifteen minutes given for each puzzle, that puzzle was scored as having been completed at the fifteen minute mark. The frequencies of participants running out of time were not significantly different between groups for any of the four puzzles; varying between $31 - 38\%$ in the first puzzle, to $9 - 14\%$ in the fourth puzzle.

Participants who received tutoring from the physical robot performed better, on average, on the second, third, and fourth puzzles than participants in any other group. Means and standard deviations for each puzzle for each group can be found in **Table 2.1** below, a plot of which is in **Figure 2.6a**. In the fourth puzzle, the mean puzzle-solving time for participants who received physical robot tutoring $(M = 7.6, SD = 3.1)$ was significantly better than the mean in either the on-screen tutoring group $(M = 8.7, SD = 2.4)$, $t(36) = 0.03$ or in the voice-only tutoring group $(M = 9.1, SD = 3.0)$ as well, $t(37) <$



(A) Mean solving time per puzzle in minutes. Participants in the physical robot tutoring condition solved the fourth puzzle significantly faster than participants in either the on-screen and voice-only tutoring conditions ($p < 0.03$). See **Table 2.1** for means and standard deviations.

(B) Mean improvement in solving time between puzzles #1 and #4. These puzzles consisted of the same gameboard, disguised in the fourth puzzle by a 90° rotation. Participants in the physical robot tutoring condition improved their solving times significantly more than participants in the other two conditions ($p < 0.01$).

FIGURE 2.6: Behavioral measure results. Participants in the physical robot tutoring condition solved the last puzzle significantly faster, and improved their same-puzzle solving time significantly more, than participants in either of the other groups.

0.02. There was no significant difference between the performance of participants who received on-screen tutoring and participants who received voice-only tutoring, across all four puzzles. This result indicates that the physical presence of the robot tutor had an effect on participants that resulted in a significant learning impact over participants who received on-screen or voice-only tutoring.

In this experiment the first and fourth puzzles were 90° rotated variations of the same gameboard. Thus, both puzzles required the same skills to solve and the difference in solving time between these two puzzles is a measure of each participant's acquired

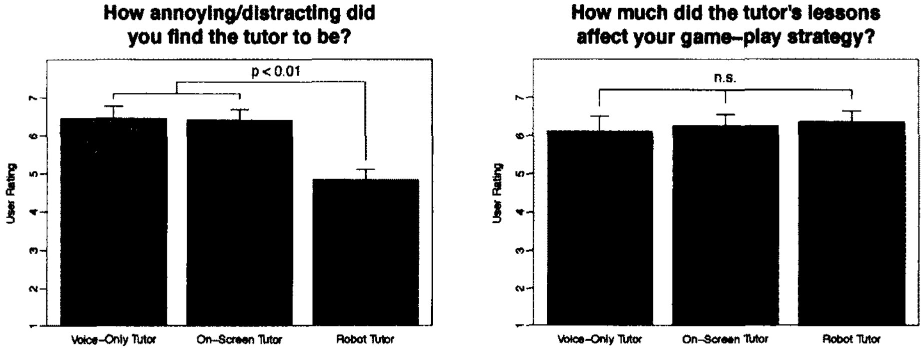
	Puzzle 1	**Puzzle 2**	**Puzzle 3**	**Puzzle 4**
Voice-Only Tutor	12.6 ± 2.4	10.7 ± 2.7	10.3 ± 3.3	9.1 ± 3.0
On-Screen Tutor	12.8 ± 2.1	11.1 ± 2.6	9.9 ± 2.6	8.7 ± 2.4
Robot Tutor	12.7 ± 2.6	10.0 ± 3.5	9.4 ± 3.0	7.6 ± 3.1

TABLE 2.1: Mean solving time across conditions, in minutes.

knowledge over the course of the study. Participants who received physical robot tutoring improved their same-puzzle solving time ($M = 5.8, SD = 3.5$) significantly more than those who receieved on-screen tutoring ($M = 3.9, SD = 2.3$), $t(31) < 0.05$ or voice-only tutoring ($M = 3.4, SD = 3.5$), $t(37) = 0.04$. There was no significant difference between the on-screen tutoring and voice-only tutoring conditions. A plot of these data can be found in **Figure 2.6b**. This result indicates that participants who received lessons from the physical robot learned more effectively than those who received voice-only or on-screen lessons.

The survey results reveal that participants found the physical robot tutor less "annoying/distracting" on average ($M = 4.9, SD = 1.2$) than participants in the other two groups, the on-screen tutor,($M = 6.4, SD = 0.8$), $t(33) < 0.05$, and the voice-only tutor, ($M = 6.4, SD = 0.7$), $t(29) < 0.05$. A plot of these data can be found in **Figure 2.7a**. This result indicates that the participants were less bothered by a physically embodied tutor.

Though the data show that the participants in the physical robot tutoring group learned



(A) Participants who received physical robot tutoring rated the tutor as significantly less annoying the the participants in the other two tutoring conditions ($p < 0.01$).

(B) Despite puzzle-solving data to the contrary, participants in all three groups rated the effect of the tutoring on their gameplay as not significantly different from one another.

FIGURE 2.7: Results of self-report measures completed after the interaction. The remaining three questions showed no significant differences between conditions.

more by the end of the experiment, those same participants did not rate the usefulness of the tutor's instruction higher than participants in the other two groups, who learned less. Responding to the survey question, "How much did the tutor's lessons affect your game-play strategy?" there was no significant difference between any group, the physical robot tutoring ($M = 6.4, SD = 0.5$), the on-screen tutoring ($M = 6.3, SD = 0.4$), or the voice-only tutoring condition ($M = 6.2, SD = 0.5$), see **Figure 2.7b**. These data indicate that whatever social effect physical embodiment has on this interaction, it did not influence the participants' perception of the value of the robot tutor over the other two tutoring conditions, despite the fact that the behavioral measure indicates better learning in the robot tutoring group.

2.10 Discussion

Perhaps the most interesting question raised by these results is: "How did the physical presence of the robot tutor improve learning gains?" The survey results do not provide a definitive answer. Participants did not report having significantly more difficulty understanding the lessons in any of conditions. In fact, all three groups rated their level of understanding of the lessons fairly highly: ranging from a low of 5.0 ($M = 5.0, SD = 1.4$) by participants in the voice-only tutoring condition to a high of 5.6 ($M = 5.6, SD = 1.2$) by participants in the robot condition. These ratings indicate that the manipulation in this experiment did not cause participants to perceive themselves as understanding more or less of the lessons as a result of embodiment. However, the performance data indicates that to some extent, they did.

Perhaps a more revealing result is the survey question that asked participants how "annoying/distracting" they found the tutor to be. Results were generally high as the survey data revealed that, in the words of one participant, "it was distracting to have the lessons interfere with my thought process unexpectedly." Participants in the physical robot tutoring condition were less annoyed ($M = 4.9, SD = 1.2$) than participants in the other two groups: the on-screen tutor, ($M = 6.4, SD = 0.8$), $t(33) < 0.05$, and the voice-only tutor, ($M = 6.4, SD = 0.7$), $t(33) < 0.05$. Perhaps this lack of "annoyance/distraction" indicates a level of respect for the physical robot that was not present in the other tutoring conditions. Perhaps participants can more easily ignore on-screen characters or disembodied voices than they can a real, physical entity.

Another hypothesis is that the learning gains can be accounted for, in part, to a social pressure to comply with the commands of a physically embodied robot, such as the effect seen in Bainbridge et al. (2008). Perhaps its physical form brings the robot closer to peer-like behavior in the subconscious minds of participants. More work is needed to understand the underlying mechanisms of this phenomena.

Another question our work raises is, "What is the duration of this effect?" Would the novelty of having a robot as a tutor wear off or does physical embodiment lead to sustainable learning gains and pedagogical advantages? A longitudinal study is needed.

2.11 Conclusion

This study investigates the effect of the embodiment of an automated tutor on adults performing a cognitively-challenging learning task. Participants who received lessons from a physically present robot tutor outperformed participants who received the same lessons from an on-screen video representation of that robot, as well as participants who received the same lessons from a voice-only tutor. Participants in the physical robot tutoring condition solved the final puzzle significantly faster and improved their same-puzzle solving time significantly more than participants in the other two groups. From these data we conclude that the physical embodiment of a tutor can yield learning gains in automated tutoring interactions.

Chapter 3

How Do Humans Personalize Their Tutoring to Students Who Tend to be More Successful vs. Less Successful?

In this chapter, we investigate human tutoring personalization in order to inform our design of human-like automated personalization systems in later chapters. In order to investigate the nuances of human tutoring personalization, we use robots as students rather than humans as students because, unlike human students, robots can be expected to perform in exactly the same way in multiple instances and with different human tutors. Studies of human-human tutoring are limited by the "chemistry" between tutor and student, which determines how effectively they are able to communicate (**Topping and Ehly 1998**). This potential confounding variable prevents human-human tutoring research from probing the nuances of human tutoring behavior. In our work, we find commonalities between how human tutors naively teach robot students and we use these commonalities to derive design guidelines for future work in automated personalization systems.

We investigate how human tutors personalize their tutoring towards students of differing histories of success in learning tasks by conducting an experiment in which each participant interacted with two robot students, one that is more successful, an "overachieving student," and one that is less successful, an "underachieving student." We measured the quantity, timing, and affective content of the instructional vocalizations that participants made towards these two robot students. We find that participants produced more speech, and more affective speech to underachieving students than to overachieving students. These results tell us that human tutors personalize their instruction based solely on the successfulness of a student and that automated systems that intend to be more human-like should treating differently-performing students significantly differently, something that is not currently done in many systems. We provide guidelines based on our findings for automated personalization systems in robot tutoring.[1]

[1] The work in this chapter was co-first authored by the present author and Elizabeth Seon-wha Kim (**Kim et al. 2009**). It appears in both authors' dissertations.

3.1 Background

In automated tutoring research, the most common forms of which are Intelligent Tutoring Systems (ITS's), most systems are designed to adapt to individual students' strengths and weaknesses (Nkambou, Bourdeau and Psyché 2010). We discuss several kinds of ITS's and their personalization systems in Chapter 1, Section 1.3. However, ITS's typically do not vary the quantity or affective content of their instruction based on the abilities of the student. Though some ITS's do model affect, they typically model the affective state of students, such as in the work of D'Mello (2012), Conati and Maclaren (2009), and D'Mello et al. (2005), rather than producing a model of affect for the automated tutor and personalizing that affect to best suit the needs individual students. The design implications that we glean here about how human tutors behave can be applied to ITS research as well as our own field of robot tutoring.

In human-human tutoring, the details of how tutors personalize their instruction to suit students of differing abilities are not fully understood. What we do know is that human tutors give individualized scaffolded guidance to students as they solve problems or analyze new concepts by providing each student with enough support to build a bridge between the student's knowledge and the content of the problem and then iteratively taking pieces of support away until students are able to build that bridge for themselves (Wood, Bruner and Ross 1976). We also know that human tutors gauge a student's understanding on an individual basis and build a mental model of a student's comprehension, which they then use to frame future scaffolding episodes (Chi et al. 2001). We discuss these features of human tutoring in more detail in Chapter 1, Section 1.2. There is no human-human tutoring work that specifically investigates whether a human tutor's verbalizations change, in quantity or affective quality, in response to the ability level of the student. We use robots as students to investigate this question.

The use of robots as students is a common practice in Learning from Demonstration (LfD) robotics research, an overview of which can be found in **Argall et al. (2009)**. The goal of LfD is to create automated systems that correctly interpret naive human teaching practices such that non-technical users can teach robots to perform novel or inherently collaborative tasks without needing to know how to program a computer. As a related topic, some LfD groups study how changing the robot student's behavior affects the kind of instruction a human tutor provides. The area this is most common is in Active Learning from Demonstration research, in which the robot student queries the human tutor for specific information about a demonstration or for specific new demonstrations **Thomaz, Hoffman and Breazeal (2006)**. This community has investigated what kinds of queries human tutors prefer to answer from a robot student and how the queries that a robot student makes affect the perception of the intelligence of that robot (**Cakmak and Thomaz 2012**). Other work in this area has found that human tutors give both instructional and motivational feedback to robot students, as well adapting their teaching strategies as they develop a mental model of how the robot student learns, all of which human tutors have also been shown to do with human students (**Thomaz and Breazeal 2008**). No work in LfD, however, has investigated how human tutors personalize their teaching to robot students of differing skill levels.

3.2 Methodology

We conducted a study to investigate how human tutors personalize their instruction when they teach robot students of differing abilities. Each participant in this study tutored two robot students, first teaching one and then teaching the other. One of the robot students was significantly more successful in the learning tasks than the other robot student. Participants were led to believe that the robots were learning based on the verbal instruction they gave, but in fact the actions of each of the robots was planned ahead of time and constant across all participants. This manipulation allows us to compare how human tutors taught these two kinds of robot students differently. We measured the quantity, timing, and affective quality of the participants vocalizations and compare how participants personalized their instruction between the more successful robot student and the less successful robot student. We use these results to inform design guidelines for future work in personalization of automated tutoring.

3.2.1 Robot

The robot we used for this study, a commercial toy called called "Pleo," is an 8-inch tall, 21-inch long green dinosaur-shaped robot created by now-defunct toy company called UGOBE Life Forms (UGOBE 2008). The robot is pictured in **Figure 3.1**. In this experiment, we used two Pleo robots, one which we called "Fred," which was the more successful robot student, and the other which we called "Kevin," which was the less successful robot student. Fred and Kevin were differentiated with different colored hats as well as separate sets of 'bark' and 'growl' vocalization recordings in order to cast

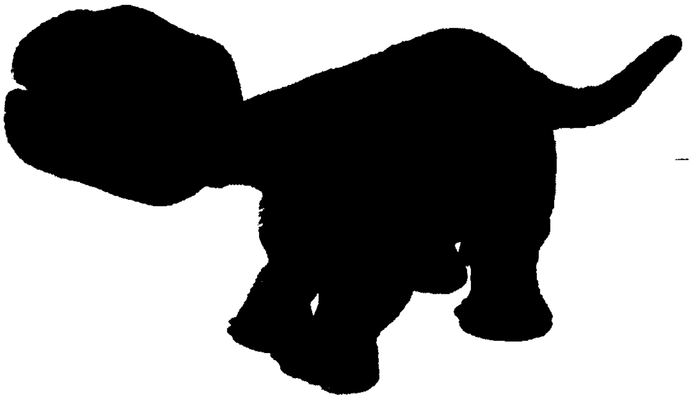


FIGURE 3.1: The "Pleo" robot, an 8-inch tall, 21-inch long dinosaur-shaped robot originally sold as a toy by UGOBE Life Forms (UGOBE 2008).

them as independent social actors in the minds of participants (Nass, Steuer and Tauber 1994).

3.2.2 Apparatus

In this study, we asked participants to teach dinosaur robots to demolish toy buildings. We used three pairs of toy buildings on each side of a model road, set up mirror image to one another, except that one building in each pair was marked with large red "X" marks. This setup can be seen in Figure 3.2. Participants taught the robot dinosaurs to knock down the buildings with red "X's," and not the unmarked buildings. The robot walked down the road towards the participant and it knocked down one of each of the pairs of buildings by first pointing to it with its head and then making a loud growling

FIGURE 3.2: Participant gives feedback to the robot student as it decides which building to knock over with its head, the building on the right or the building on the left.

noise and swinging its head forcefully into the building in order to knock it over. We asked participants to speak to the robot to guide it through this demolition process.

We conducted the study on a 10-inch wide by 30-inch long model road along which the robot walked straight across toward the participant. On each side of the road, there were three cardboard toy buildings, approximately 10-inches tall. The buildings on each side of the road were pairwise identical such that the left side of the road mirrored the right side of the road, except for the red "X" marks. First, the robot encountered a pair of

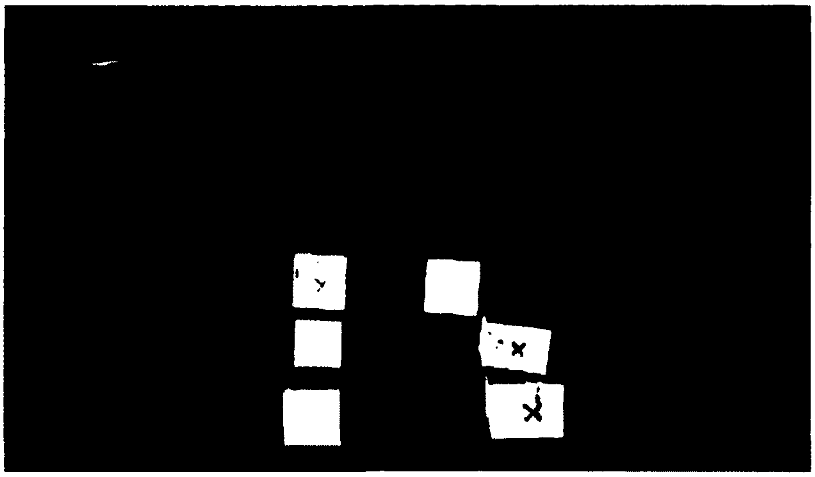


FIGURE 3.3: The overhead view used for Wizard of Oz control of the robot's locomotion. North of this frame, a participant is standing at the end of the table. Building pairs, from bottom (beginning) to top (end) are: purple, silver, and orange.

purple buildings, then a pair of silver buildings, last a pair of orange buildings, which were the buildings closest to the participant. This setup is pictured in **Figure 3.3**.

The road on which this task took place was set on a table about 3 feet off the ground. The three pairs of buildings were placed on either side of a straight, yellow double-lined road. The yellow double-lines were raised, providing a track for the robots to walk along, ensuring that the robot stayed in the middle of the road at all times. The buildings were separated from each other on each side of the road by a space of 3 inches. From the perspective of the robot, the buildings that were marked with the red "X's" were: the purple building on the right side of the road, the silver building on the right side of the road, and the orange building on the left side of the road, as seen in **Figure 3.3**. These markings and orderings were constant for all participants.

3.2.3 Conditions

There were two conditions in this study and each participant saw both conditions.

One condition, in which the robot was called "Fred," was the more successful of the two students. The other condition, in which the robot was called "Kevin," was the less successful of the two students. The only difference between the behavior of Fred and Kevin is that Fred always chose the correct building to knock down for all three pairs of buildings, whereas Kevin chose the incorrect building in the first and second pairs, but chose the correct building in the last pair.

The ordering of Fred-then-Kevin, or Kevin-then-Fred, was alternated per participant. Of the 27 total participants, 13 saw Kevin first and 14 saw Fred first. We investigate whether the ordering of the two robots affected the participants vocalizations with two-way ANOVAs in the results.

3.2.4 "Wizard of Oz"

It was essential for the success of this study to convince participants that the robot tutor was listening and responding quickly and accurately to their instructions. Automated Speech Recognition (SR) and automated Affect Recognition (AR) systems are not yet as robust or reliable as human speech and human affect recognition, see **Gold, Morgan and Ellis (2011)** and (**Zeng et al. 2009**) respectively. Therefore, in order to guarantee the perception of human-like responsiveness, the robots in this study were secretly controlled by a remote operator. This is an experimental methodology called "Wizard of Oz"

(Dahlbäck, Jönsson and Ahrenberg 1993) which allows us to create the illusion that the robot can autonomously respond to the content of the participants' tutoring as some day SR and AR systems will enable it to do.

The "wizard" responded to the participants' vocalizations with a set of seven pre-defined actions: a happy bark sound, a questioning bark sound, a sad bark sound, a move forward, swinging its head left, swinging its head right, or doing a "happy dance" at the conclusion of the study.

To give the appearance of autonomy, the robot also had several idling behaviors that were not controlled by the wizard. In the event that no action was taken by the wizard in three seconds, the robot would make a yawning sound, or a quiet huffing sound, to indicate idling. It would accompany those sounds with slight head tilts and shifting of its legs, in random order, again to give the illusion of lifelike autonomy.

The wizard was off-site and never met or interacted with the participant during the course of the study. There was no indication that participants were aware of the presence of a wizard in any of the trials. The wizard was able to hear the participant through a clip-on lapel microphone we asked all participants to wear. The wizard was able to see the participant via two live camera feeds on a television screen and a laptop showing the same perspectives as seen in Figure 3.2 and Figure 3.3.

The robots were controlled by infrared (IR) signals. There is an IR receiver in the nose of each robot. IR signals were sent from long-distance IR beacons through an IguanaIR USB-IR transceiver (IguanaWorks 2008), controlled in Linux using LIRC (Linux Infrared

Remote Control) software (Bartelmus 2008). The beacons were located in front of the participant, disguised as an additional camera. The wizard controlled the robot using the seven pre-scripted/pre-recorded behaviors described above with a handheld USB gaming pad, the buttons of which were mapped to each behavior. For example, pressing the forward button caused the robot dinosaur to walk forward, and pressing to the left or right caused the robot to swing its head in the respective direction. These robot actions were created and modified using UGOBE's software development kit and a third-party Pleo development platform called MySkit (DogsBody & Ratchet Software 2009).

The "Wizard of Oz" methodology was used here to ensure participants were convinced that the robot could hear and react to their instructions quickly. This is essential to the success of this study because if participants ever doubted the ability of the robot to respond to their instruction, they would have been disincentivized from providing any further instruction and that would have adversely impacted our data collection and results. In this study, we are investigating the effect of the successfulness of the robot tutor, which was held constant between participants based on each of the two conditions. The human operator of Fred and Kevin followed the exact same protocol for the learning tasks across all participants. Thus, we can compare how any individual participant may have treated Kevin differently from Fred based on their relative successfulness as students.

3.2.5 Participants

There were 27 participants in this study, 9 male and 16 female, each 18 years of age and above. Our exclusion criteria were a lack of English fluency or previous research or coursework experience in artificial intelligence or robotics.

3.2.6 Procedure

The testing session lasted approximately 30 minutes. Each participant gave informed consent to be recorded, and then was led into a lab containing the two dinosaur robots and the road and building apparatus described above. The participant stood behind the end of a table and clipped a lapel microphone to his/her shirt collar. Fred and Kevin, the robots, stood in front of the demolition training course, close to and facing the participant.

The participant was told the following:

"These are our dinosaurs, their names are Kevin and Fred. Kevin is the one with the red hat with the 'K' on it. Fred is wearing a blue hat with the letter 'F'. Today they're going to train to join a demolition crew. They'll be knocking over buildings with their heads. Behind them is the training course that they'll running today. They'll go one at a time: Fred will be first and I'll take Kevin and leave the room. When Fred's done, then it'll be Kevin's turn."

The ordering of the dinosaurs varied per participant as specified above, name orderings were changed accordingly in the instructions.

The participants were then told, *"You are going to help them pick the red 'X'-marked buildings in the training course to demolish. In the training course, you'll see there are three pairs of colored buildings standing across from one another – the purple pair at the far end, the silver pair in the middle, and the orange pair closest to us. The robots will do the training course sequentially, starting at the purple buildings and walking towards us. For each pair, you'll see that one is marked with an 'X.' Kevin and Fred can see the 'X's too. For each pair of buildings it's important that they knock down the building with the 'X' and that they don't knock down the unmarked building."*

"They already know how to knock down buildings. We want you to help them understand that they should only knock down the buildings with the red 'X's and all of the ones with the 'X's. You're going to help them by talking with them. We encourage you not to make any assumptions about how this might work. Just act naturally and do what feels comfortable. Please stay in this area demarcated by caution tape. The training is complete when an orange building falls."

The experimenter then asked the participant to say hello and explain the task to the robots, in his or her own words. The dinosaurs returned the greetings with a happy bark and acknowledged the receipt of instructions with another happy bark closely following the participant's utterances. The experimenter then solicited questions or provided additional clarification for the task from the participant.

Then the experimenter placed one of the dinosaurs at the start position, between the first pair of buildings at the far end of the course, facing the participant. The experimenter left the room, taking the other robot with them out of the room. The participant was then alone with the robot.

The robot gave a happy bark vocalization indicating the start of the trial. The robot then walked to the first pair of buildings. Then it slowly (over 4 seconds) communicated its intent to knock over one of the first pair of buildings, by turning his head towards the building while vocalizing a slowly increasing growl. If the participant did not vocalize negatively towards the robot, the robot concluded his swing into the building and the building fell. If the participant did say "stop," or a similar instructive command, the dinosaur discontinued the swing towards the originally intended building, and turned its head towards the other building and again began vocalizing its intention to knock down the other building by swinging its head towards it slowly. After one of the first pair of buildings fell, the dinosaur walked forward to the second pair of buildings and repeated this procedure. After finishing with the second pair, this procedure repeated again for the third pair of buildings.

The experimenter returned to the training room when either the participant indicated the end of the training, or a period of time (approximately 30 seconds) elapsed after one of the last pair of buildings fell. The participant was given a few minutes' break while the experimenter reset the demolition training course, putting all buildings right-side-up again. The participant then engaged in this same procedure with the remaining robot.

The only difference between these two training sessions was the original intended choice of the two robot learners in each of the three trials. The robot named "Fred" always chose the correct building, in all three pairs. The robot named "Kevin" chose the incorrect building in the first two pairs but the correct building among the third pair.

Once the second training session was complete, the participant took an exit survey. Afterwards, the experimenter debriefed the participant and showed him or her the "Wizard of Oz" control room, explained the technology, the purpose of the study, and the necessity for the deception.

3.3 Analysis

After conducting the experiment, we divided the recordings of participants into three phases and we used human coders to analyze the recordings to assess their affective content.

3.3.1 Vocalization Categories

We noted participants' vocalizations fell into three cyclic sequential phases, based on the robot's progress in each trial. All three of these phases occurs in each trial: (1) 'Direction,' which occurred before the robot picked a building, (2) 'Guidance,' which occurred while the robot swung its head to knock over a building, and (3) 'Feedback,' which occurred after the building fell or the robot stopped its swing. We segmented all our audio recordings into these three categories.

For example, when the first of the three trials began with the robot placed between the first pair of buildings, where the robot indicated its readiness by vocalizing. The robot then signaled its intent to knock over a building, either the one on the left or the one on the right, where the intention is broadcast for four seconds. Last, if the robot was not corrected or stopped, it knocked over the building it signaled intent to knock down. In this example, the first sentence describes the 'Direction' phase, the second describes 'Guidance' and the last is 'Feedback'.

In this manner, for each pair of buildings, the instructional phases cycled from 'Direction' to 'Guidance' to 'Feedback,' then back to 'Direction'. Sometimes there was one cycle per trial: the robot gets to the building pair, swings its head towards the correct building of the two, and knocks it down. Other times, there were two: the robot gets to the building pair, intends for the wrong building and receives reprimand, then replies to the reprimand, then intends for the right building, and knocks it down.

The segmentation was performed by recognizing the robot sounds we heard on the audio recordings that uniquely identified the phases of each trial. The only phases for which there was no unique sound indication was between trials: separating the last phase of one trial ('Feedback') and the first of the next trial ('Direction'). We waited for a two-second pause in our participants' vocalizations, but if there was none we divided based on the transcription of the words used such that 'Feedback' ended when evaluative words were no longer used, such as "no," "stop," "right," or "good job."

3.3.2 Analyzing Affective Content

We segmented the audio recordings of each participant's vocalizations according to the guidelines above. The average length of each file was approximately 20 seconds. We then randomized the ordering of these files and asked two human coders who were blind to the design and conditions in this study to identify the number of words in each file and to analyze the affective content of each file.

The coders rated the affective content of each audio clip as either positive, negative, or neither. Positive affect was described to coders as sounding "encouraging," "approving," or "pleasant," whereas negative affect was described as sounding "discouraging," "prohibiting," or "disappointing." We asked the coders to rate the intensity of the affect on a differential semantic scale originally conceived by **Osgood, Suci and Tannenbaum (1957)**, from 0 (mild) to 2 (very strong), and their respective confidences for each judgement on a differential semantic scale from 0 (not sure) to 2 (quite sure).

3.4 Results and Discussion

This study investigates how human tutors personalize their instruction towards robot students of differing abilities. The measures we used in this study were the number of words spoken per second, the affective category ratings (i.e., positive, negative, or neither), and the affect intensity ratings (from 0, "mild," to 2, "very strong"). The ratings of two naive coders showed high agreement ($\kappa = 0.84$ using Cohen's quadratically weighted, normalized test (**Cohen 1968**). Most audio clips were short and contained 7.71 words/clip on average (1.26 words/sec) and a standard deviation of 8.43 words/clip (1.36 words/sec). Because teaching styles varied per participant, where some participants were more verbose than others in communicating the same information, we performed our

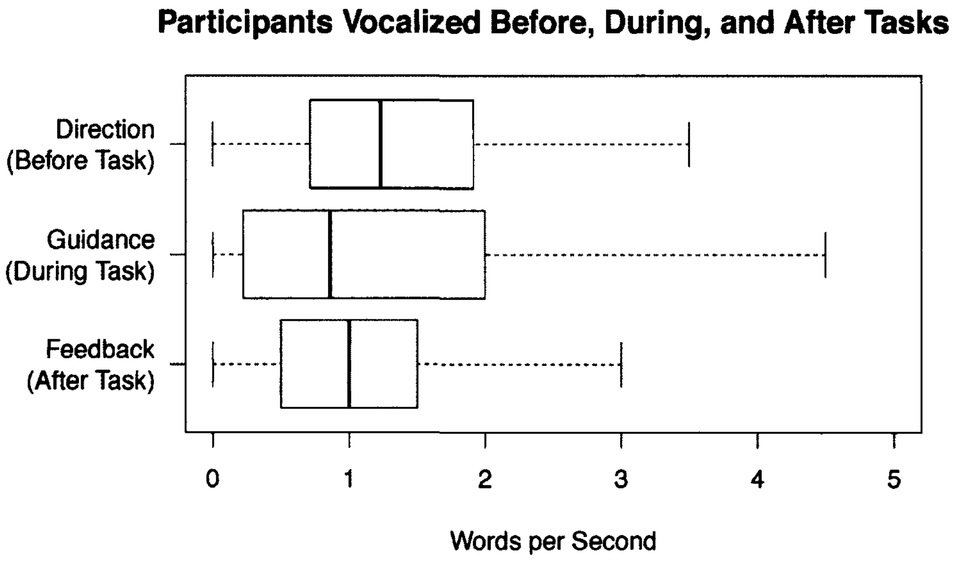


FIGURE 3.4: Distribution of vocalizations across the three phases: before, during, and after each learning task. No significant difference between phases indicate that participants provided a similar amount of instruction in each phase of tutoring. The boxes in this plot contain the middle 50% of observations, whereas the whiskers extend to the outer quartiles.

analysis using 'words per second' as our main measure, rather than 'words per clip'. Using 'words per second' allows us to compare how much of the time participants were speaking to the robot over the course of their interaction, regardless of how many words they used to communicate the instructional content.

We present our findings here and we use the findings to propose design guidelines for future work in automated personalization systems such as the ones we later design for robot tutors.

3.4.1 Human tutors vocalize before, during, and after a learner's actions

Participants in this study used an almost equal number of words per second in all three phases: before, during and after each learning trial. A plot of this data can be found in **Figure 3.4**. Over all phases, the frequency of words spoken was on average 1.26 words/sec, with a standard deviation of 1.36 words/sec. There were no significant differences between groups. This result indicates the human tutors provide the same amount of tutoring vocalizations per second before, during, and after learning tasks.

Typical Intelligent Tutoring System's (ITS's) do not provide guidance all throughout the learning tasks (**Nkambou, Bourdeau and Psyché 2010**). The two dominant families of ITS's (as outlined in **Chapter 1**, **Section 1.3**) provide either step-by-step feedback during a problem, or they provide feedback after a problem, and some hybrid systems do both. These two kinds of instructional content are most closely related to our second and third phrases, what we call the 'Guidance' and 'Feedback' phases. Our findings indicate that for automated tutors to behave more like human tutors, if operating in a similar educational context as this one, they need to provide instruction to students that includes a significant amount of planning before a task, as in our 'Direction' phase, which most automated tutoring systems do not currently do.

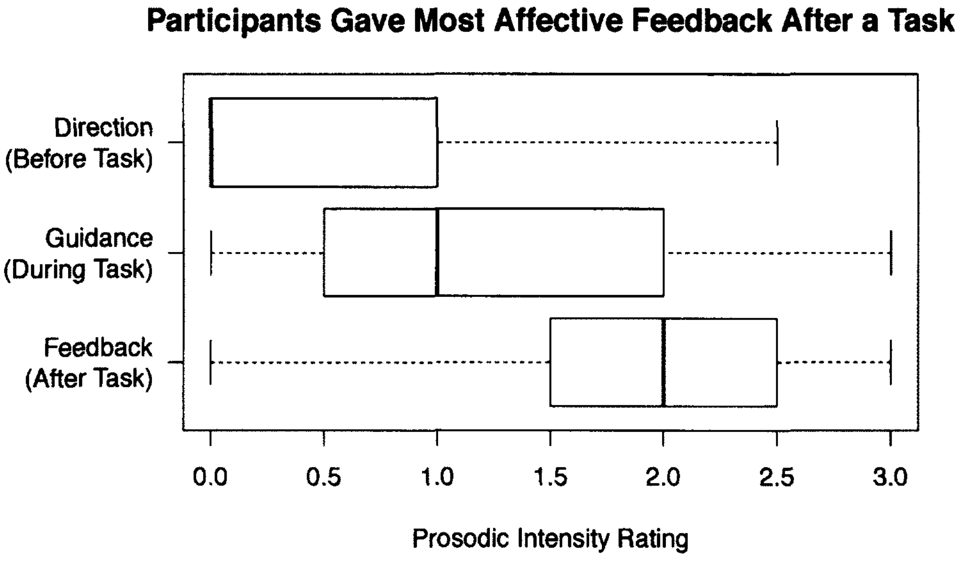


FIGURE 3.5: The affective intensity of the instruction participants gave was significantly higher in each successive phase of the tutoring task: before the task, during the task, and after the task, ($p < 0.001$ for the ANOVA tests, $F[2] = 58.2, 19.2$).

3.4.2 Human tutors express affect during and after a learner's actions

Although we did not specifically instruct participants to use affective content in their instruction to the robot tutors, participants vocalized with intensely affective prosody during 'Guidance' phase ($M = 1.28, SD = 0.93$) and in the 'Feedback' phase ($M = 1.89, SD = 0.78$). Figure 3.5 plots this data. Participants' affective intensity in the 'Direction' phase was minimal ($M = 0.47, SD = 0.68$). The differences in affective intensity ratings between 'Direction' and 'Guidance', and between 'Guidance' and 'Feedback', were both significant ($p < 0.001$ for both ANOVA tests, $F[2] = 58.2, 19.2$). This indicates that participate gave significantly more strongly affective instruction in each successive phase of the task: from before the task to during the task as well as from during the task to after the task.

This is in stark contrast to most automated tutoring systems which do not model the affective production of the tutor's dialogue. Our findings indicate that to make automated tutors more like human tutors, automated tutors should produce more strongly affective feedback during a learning task than before one, and more strongly affective feedback after a learning task than during one.

3.4.3 Human tutors help less often as a learner continually succeeds

Participants used significantly fewer words per second when teaching "Fred," the more successful robot student, over the course of the three subsequent trials ($p < 0.002$, linear regression). **Figure 3.6** is a plot of this data. We verified that this trend is not explained by the condition ordering. In a two-way ANOVA, we found a highly significant main effect for trial number ($p = 0.0018, F[1] = 10$) and for order ($p = 0.0004, F[1] = 13$), but not for their interaction ($p = 0.38, F[1] = 0.7$). These data indicate that the drop-off in words per second of instruction was not due to repetition or boredom, but rather that human tutors provide progressively less frequent instruction to a more successful student over time. A similar test for "Kevin," the less successful robot student, showed no trend of decreasing words per second over the three trials ($p = 0.57, F[1] = 0.38$).

This result indicates that for automated tutoring systems to be more like real human tutors, they should provide less frequent tutoring to students who consistently do well. Currently, most automated tutoring systems provide the same frequency of instruction to

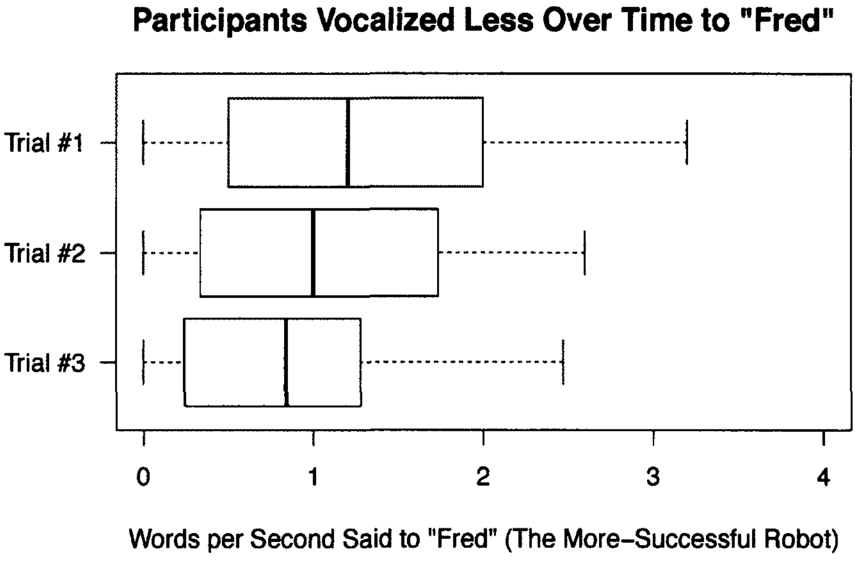


FIGURE 3.6: Word counts per second for vocalizations made to "Fred," the more successful student. Over the course of the three subsequent trials, the number of words per second decreases significantly ($p < 0.002$, linear regression).

all students, or they allow students to select more or less feedback (**Nkambou, Bourdeau and Psyché 2010**). Our results show that human tutors significantly vary the rate of instruction between students, providing more instruction to less successful students.

3.4.4 Human tutors give more instruction to a student they perceive as struggling

In the third trial for each robot, after participants had experienced the first two trials in which "Fred" picks the correct answer each time and "Kevin" picks the incorrect answer each time, participants gave significantly more guidance to the less-successful robot ($p <$

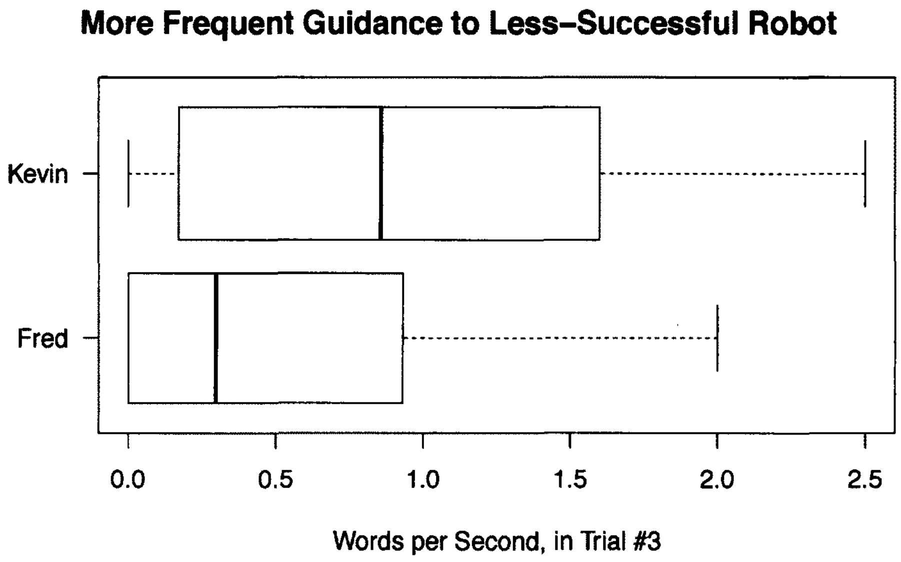


FIGURE 3.7: In the third trial, after participants had experienced the first two trials in which "Fred" picks the correct answer each time and "Kevin" picks the incorrect answer each time, participants gave significantly more frequent guidance to the less-successful robot ($p < 0.05, F[1] = 5$).

$0.05, F[1] = 5$). This result indicates that participants formed distinct mental models between Kevin and Fred, and in anticipation of another failure by the less successful student, produced more instructional content than they did for the student who continues to do each task correctly.

This result is consistent with related work in Learning from Demonstration (LfD) that has shown that human tutors build models of each individual robot learner, and personalize their instruction based on those models (**Thomaz and Breazeal 2008**). We show here that not only do they personalize their instruction but they also build expectations of which robot will need more instruction and which will need less.

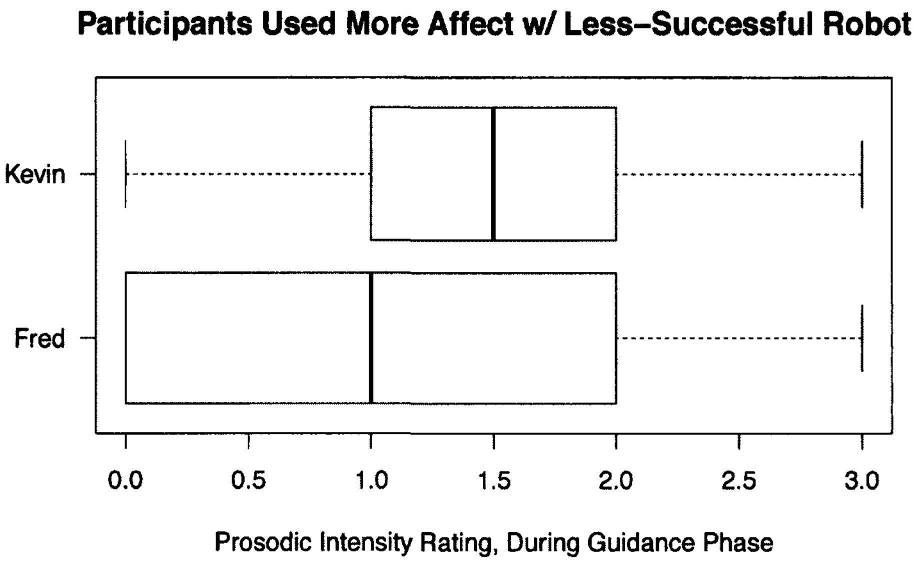


FIGURE 3.8: In the 'Guidance' phase, which is the phase during the learning task, participants gave more intensely affective instruction to the less-successful robot than they gave to the more-successful robot, ($p < 0.05, F[1] = 5$).

3.4.5 Human tutors used more intensely affective vocalizations toward the less-successful robot

In the 'Guidance' phase, which is the phase that takes place exclusively during the learning task, participants gave more intensely affective instruction to the less-successful robot than they gave to the more-successful robot, ($p < 0.05, F[1] = 5$). A plot of this data can be found in **Figure 3.8**. 83% of the affective vocalizations given to the less-successful robot in this phase were rated as positive in nature by our independent coders; 17% was rated negative. This result indicates that human tutors increase the affective content of their vocalizations by producing more positive or encouraging vocalizations when a student is struggling than when a student is succeeding.

This result contrasts with the majority of automated tutoring systems that do not employ an affective model for their instruction (**Nkambou, Bourdeau and Psyché 2010**). We show that human tutors do personalize the affective content of their instruction to students of differing abilities. For automated tutors to act more like human tutors, we must build models of affect for the dialogue that automated tutors produce.

3.5 Conclusion

In this chapter, we investigated how human tutors personalize their instruction to robot students of differing abilities. Each participant taught two robot students, one more successful in the learning tasks and one less successful in the learning tasks. We found that participants personalized their instruction between robots such that the less successful student got more instruction and more strongly affective instruction than the more successful student. We also found that participants gave less instruction over time to the more successful student. We provide design guidelines based on these findings for future work in automated personalization systems for tutoring:

- We suggest that to make automated tutors more like human tutors, automated tutors should produce more strongly affective feedback during a learning task than before one, and more strongly affective feedback after a learning task than during one.

- Our findings indicate that for automated tutors to behave more like human tutors, if operating in a similar educational context as this one, they need to provide instruction to students that includes a significant amount of planning before a task, as in our 'Direction' phase, which most automated tutoring systems do not currently do.

- We show here that not only do they personalize their instruction but they also build expectations of which robot will need more instruction and which will need less.

- We suggest that for automated tutors to act more like human tutors, we must build models of affect for the dialogue that automated tutors produce.

Chapter 4

How Do Humans Personalize Their Tutoring to Robots with Differing Emotional Responses?

In this chapter, we continue our study of human tutoring personalization by pairing human tutors with robot students. Our previous work investigating how human tutors personalize their interaction to students of *different abilities* leads us to ask how human tutors might personalize their interaction to students of the *exact same ability*, but with differing emotional response patterns. By using robot students we can investigate how humans tutors personalize their instruction to students who perform exactly the same way on a series of learning tasks but whose emotional responses differ significantly. We use our findings to derive several design guidelines for future work in automated tutoring systems.

In order to investigate how human tutors personalize their instruction to robot students that perform the same way in learning tasks but have differing emotional responses, we designed an experiment with three conditions in which participants tutored a robot student that: gave either (1) typical emotional feedback to the human tutor, (2) apathetic emotional feedback, or (3) atypical emotional feedback. In all three conditions the robot student performed the learning tasks in the exact same pre-scripted way, and it received the exact same pre-scripted grades based on its performance. We led participants to believe that their instruction helped the robot learn to perform the learning task better over time and we allowed participants to choose how many demonstrations of each lesson they would give to the robot student. We measured how many demonstrations participants chose to give as well as the precision with which the participant was performing each demonstration. We use these sources of data to investigate whether only the personality of the robot student, and not any differences in learning performance, influence human tutoring personalization. We use the results of this study to propose guidelines for creating more human-like automated personalization tutoring systems.

4.1 Introduction

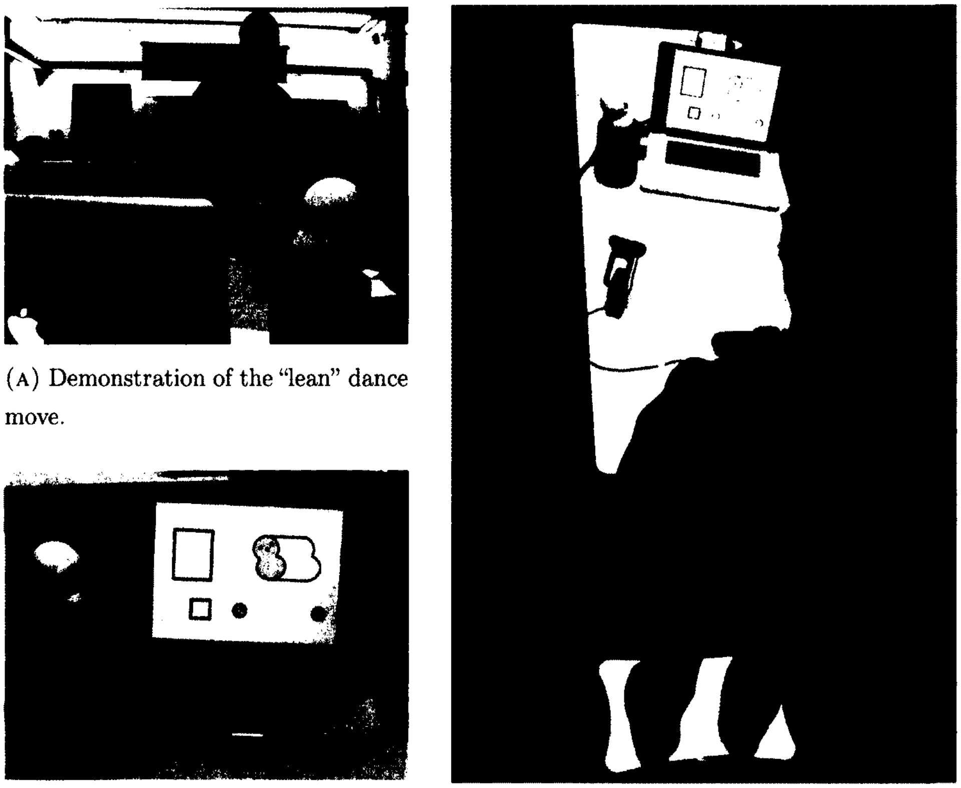
We do not know the precise underlying mechanisms by which human tutors personalize their tutoring. It has been shown that human tutors personalize their instruction based on how students perform in learning tasks (**Wood, Bruner and Ross 1976**) and it has also been shown that human tutors personalize their instruction based on student's affective state (**Lehman et al. 2008**), but it is not known how these two effects are related. How would human tutors personalize their instruction if two students performed identically across all learning tasks but had differing patterns of affective expressions? We investigate this question with robot students.

Automated tutoring systems that take into consideration the student's affective state are becoming more common in Intelligent Tutoring Systems (ITS) research (**D'Mello et al. 2005**). From the work incorporating affective sensors into AutoTutor, in which the tutor detects affective states like frustration, delight, flow, and confusion (**Craig et al. 2004**), to systems in which boredom is closely monitored (**San Pedro et al. 2013**), there are now a variety of systems that take into consideration the emotions of the student. How to personalize the tutor's instruction with this affect information is not a trivial question. In this study, we explore how humans personalize tutoring to robot students who vary their affective state but not their learning performance in order to isolate the variable of affect and inform future research in automated tutoring systems that personalize instruction based on a student's affective state.

We tested three affective state conditions in this study: either (1) emotionally appropriate responses, (2) often emotionally inappropriate responses, or (3) apathetic responses. We chose to manipulate the emotional appropriateness in these conditions, along with a control for apathetic responses, because it is known that when people are presented with inappropriate emotional expression in other humans, they question their own perception of events in an attempt to 'correct' the inconsistency, a phenomenon known as "cognitive dissonance" (Aronson 1969). Related work that compared participants' impressions of a virtual agent, outside the education domain, where the agent either was or was not consistent in its emotional expression replicated the findings about human-human interactions in human-agent interactions (Creed and Beale 2008). We investigate here whether cognitive dissonance can affect a human tutor's personalization, even when students perform in exactly the same way otherwise. The cognitive dissonance effect allows us to definitively answer whether or not emotion alone affects human tutoring personalization.

4.2 Methodology

In this study we investigate how human tutors personalize their instruction to students who perform the same but having differing emotional responses to that performance. For this experiment participants were asked to teach the robot several "dances" by demonstrating them repeatedly for the robot student. During each dancing demonstration, the robot would dance as well and, after each demonstration, the robot would receive a score based on its performance. The robot would then respond to that score. Across all three conditions in this study, the robot danced in exactly the same way and received exactly the same pre-programmed scores. The only differences between the three conditions were the emotional statements the robot student made after receiving those scores: they were either (1) emotionally appropriate responses, (2) often emotionally inappropriate



(A) Demonstration of the "lean" dance move.

(B) The participant's view of the apparatus: the robot and the dance instructions on screen behind the robot.

(C) The apparatus viewed from above. Participants stood on the Wii Fit Balance Board, visible at the bottom of this image.

FIGURE 4.1: The experimental apparatus. Participants were asked to demonstrate dances to a robot, where the instructions for the dances were displayed on the screen behind the robot. We led participants to believe that the robot learned dances by watching the participant's demonstrations.

responses, or (3) apathetic responses. Participants saw exactly one of these conditions and we allowed participants to choose the number of demonstrations to do for each dance. We measured the number of demonstrations participants choose to do as well as the accuracy of those demonstrations in order to study the effect of this manipulation of emotional responses on human tutoring personalization.

4.2.1 Apparatus

Participants were asked to demonstrate five predefined "dances," each set to a unique half-minute segment from popular music. Table 4.1 lists the five song segments we chose. We choreographed a unique "dance" for each of the five song clips, of progressively increasing difficulty, the details of which are described below. Participants were led to believe that demonstrating the dances for the robot student would teach the robot to perform the dances. Unbeknownst to the participants, the robot student performed a pre-determined sequence of dances, with built-in failures, exactly the same way for each participant regardless of the participant's input. We found only one participant who caught on to this manipulation and his data was excluded from the results.

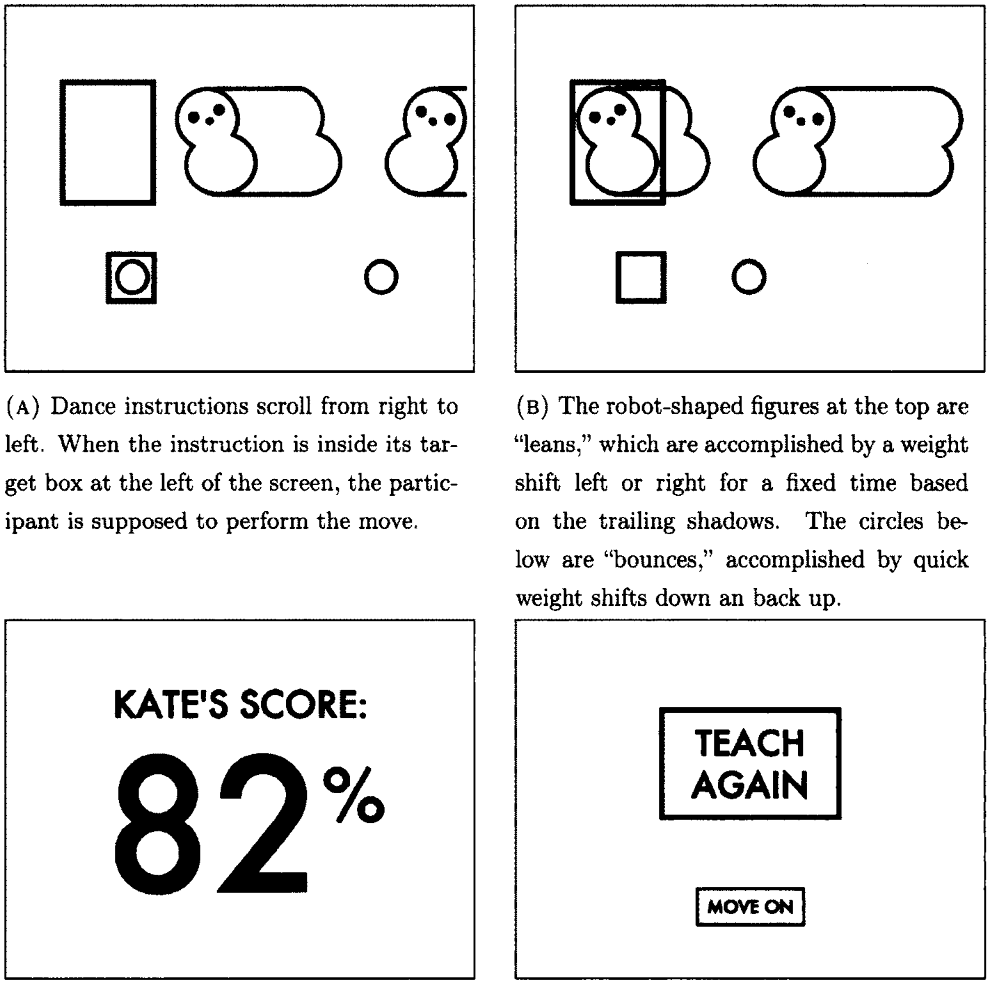
#	**Artist**	**Title**	**Cut**
1	Willy Wonka	Oompa Loompa	$0:20-0:51$
2	Daft Punk	Robot Rock	$0:34-1:02$
3	Michael Jackson	Billy Jean	$0:26-0:58$
4	Basement Jaxx	Do Your Thing	$0:32-0:59$
5	Lady Gaga	Just Dance	$0:46-1:22$

TABLE 4.1: The song clips that were used in this study.

4.2.2 "Dances"

Throughout the experiment, participants stood on a Nintendo Wii Fit Balance Board Peripheral, which is a wide and low-to-the-ground pressure-sensitive platform (**Nintendo 2007*b***), placed in front of the robot. This peripheral is pictured in **Figure 4.1c**. Participants were given dance instructions on a screen behind the robot, as seen in **Figure 4.1b**. These dance instructions were positioned behind the robot and out of its line of sight. Instead, as participants followed the dance instructions, we led them to believe the robot was mimicking their movements by having the robot perform movements close to those in the instructions participants were following. The robot appeared to mimic participants during every demonstration. After each of these demonstrations, the robot would turn to face the computer and it would receive a score out of 100, as seen in **Figure 4.2c**. The reaction the robot gave to this score was the only difference between conditions in this study.

The "dances" themselves were composed of series of two kinds of moves: (1) 'leans', either left or right, and (2) 'bounces'. To perform a lean, the participant would shift his or her weight to one side of his or her body. Leans had varying durations, indicated by a trailing shadow of the robot image on the screen, as seen in **Figure 4.2b**. The 'bounce' move was performed by bending one's knees and then quickly standing upright again. Bounces did not have varying durations, instead they were intended to be performed as quickly as possible. Bounces could be executed during leans, or on their own. On average, there were 13 seconds of leaning and 16 bounces per 30-seconds of dance. The



(A) Dance instructions scroll from right to left. When the instruction is inside its target box at the left of the screen, the participant is supposed to perform the move.

(B) The robot-shaped figures at the top are "leans," which are accomplished by a weight shift left or right for a fixed time based on the trailing shadows. The circles below are "bounces," accomplished by quick weight shifts down an back up.

(C) After each demonstration, the robot receives a percentile score, turns around to look at it, and responds with one of three kinds of verbal responses, depending on the experimental condition.

(D) After every dance, the participant chooses whether to demonstrate that dance again or move to on to the next dance.

FIGURE 4.2: Screenshots of the user interface.

dances ranged in complexity from 8 to 30 bounces per dance and from 8 to 20 seconds of cumulative leaning per dance.

The dance instructions were given in an illustrated scrolling interface style similar to the interfaces found in popular rhythm-based video games like *Dance Dance Revolution* (**Konami 1998**) and *Guitar Hero* (**Harmonix 2005**). Our interface is depicted in **Figure 4.2a** and **Figure 4.2b**. In the interface, there are robot-shaped figures representing the lean dance moves, and circle figures representing the bounce dance moves. These figures start at the right-hand side of the screen and scroll slowly towards the left-hand side. On the left-hand side are two stationary targets, in the shape of black rectangles. When the moving figures, starting on the right, reach the stationary targets on the left, participants were told to perform the dance moves they illustrated. In this way, participants could monitor the fixed targets for the dance moves they should do at the current moment, and they could look towards the right of the targets to see the dance instructions coming up next. The lean dance moves lingered in the fixed target for the length of time it took for the trailing shadow to catch up with the target – such that longer shadows indicated longer leans and shorter shadows indicated shorter leans. Bounce dance moves lingered in the target for half of a second each.

We chose these two dance moves, "leans" and "bounces," in order to best utilize the kind of data that Wii Fit Balance Board Peripheral provides, which is four weights representing the force applied to each quadrant of the board. This allows changes in position, especially along the X and Y axes, to be easily detected. "Leans" and "bounces," X and Y axis shifts respectively, were used because they were easy for participants to do, easy to display on screen, and easy for the robot to mimic. We calculated accuracy scores for each participant for each demonstration by an average of two values: (1) the percentage of time that a participant was kneeling down during the approximately half of a second of the "bounce" symbol stopping in its target rectangle and (2) the percentage of time that the participant leaned his or her weight in the direction of the "lean" indicated in its target rectangle.

4.2.3 Robot

The robot we used for this study is the same robot we used in **Chapter 2** to study the effects of embodiment on robot tutoring. The robot, "Keepon," is a small, yellow, snowman-shaped device with four degrees of freedom. For a description of its capabilities, see **Chapter 2**, **Section 2.5.1**. The robot was referred to as 'Kate' throughout this experiment.

During the course of the experiment, when the robot was not dancing, it looked around the room at randomly chosen degrees of rotation and occasionally made humming noises, breathing noises, sighs, or yawns. These idling behaviors were intended to cajole the participant into making a choice on the screen so as to start or continue the experiment. In addition, the robot confirmed selections made by the participant on the screen by speaking phrases like "Oh, okay, let's move on!" when the participant chose to move on to the next song, or "Here we go!" or "Okay. I'm ready!" when he or she chose to begin demonstrating a dance. Lastly, during the dance itself, the robot occasionally spoke one of several 'thinking' sounds, like "Hmm." or "Oh!" These additional speech utterances were timed at random, intended to give the illusion of the robot's autonomy.

During each demonstration, the robot also danced. The robot's movements corresponded to the dance instructions displayed on screen, in proportion to the score that it received for that demonstration. For example, to achieve a score of 78%, the robot would perform only 78% of the moves indicated in the instructions. For the other 22%, the robot would remain motionless. We intended for this lack of motion, in addition to the "thinking sounds" described above, to communicate that the robot was watching the participant's demonstration in the time that it was not performing the dance itself.

4.2.4 Scores

The scores the robot received were percentages proportional to the the robot's dancing accuracy compared to the on-screen instructions. Unbeknownst to the participants, the sequence of scores (and the performance of the robot) was pre-scripted for each demonstration of each dance. For example, on the third demonstration of the third dance, the robot received a score of 80% regardless of how well the participant demonstrated the dance and regardless of what experimental group he or she was in. On the next demonstration, the fourth, the robot would always receive a score of 84%. With every subsequent demonstration of a dance, the score increased. This was done across all conditions, in order to isolate the effect of the responses to these scores by holding constant the successfulness of the student .

Each dance had a separate sequence of scores, prepared in advance, but all of the sequences began with several low scores (all below 30%), followed by a large jump to a series of higher scores (all above 75%). The jump occurred on the third demonstration for each of the first three dances, on the fourth demonstration of the fourth dance, and on the fifth demonstration of the fifth dance. The intention of these jumps in the scores was to provide participants a convenient stopping point for each dance. We investigate how many participants in each demonstration were patient enough with the robot student to reach the jump in each of the dances pre-planned scores.

4.2.5 Conditions

What the robot said in response to the scores it received was the only difference between the three conditions in this study. These responses contained between two and fifteen spoken English words, all recorded in the same female voice. Sample responses for all three conditions can be found in **Table 4.3**, **Table 4.4**, and **Table 4.5**. Participants were exposed to an average of 3.8 to 5.9 robot responses per song, depending on how many demonstrations they elected to perform.

We base the emotionally appropriate responses condition and the often emotionally inappropriate responses condition on a subset of two of the appraisal dimensions defined by the EMA model of emotion (**Marsella and Gratch 2009**). The two dimensions we used were:

- **Desirability**, which reflects robot student's appraisal of the scores it earned, where scores above 75% was considered desirable and scores below 30% were considered undesirable. All scores in this study were set to be either below 30% or above 75%. The motivation for this choice is documented below.

- **Expectedness**, the robot's expectation of a score based on its previous score. The first score for each dance was always treated as unexpected. After the first score, when the robot's score changed by 10% or more from one demonstration to the next, it was treated as unexpected. In all other cases, the score was treated as expected.

We treat both appraisal dimensions as binary decisions, yielding four possible emotional categories to describe the "emotionally appropriate" response for any given score. A description of these four categories is found in **Table 4.2** below.

	Expected	**Unexpected**
Desirable	satisfaction, pride	happy-surprise, relief
Undesirable	shame, frustration	disappointment, worry

TABLE 4.2: The four emotional categories from which we determine the "emotionally appropriate" response.

For each of the four categories, we recorded approximately fifteen spoken emotional utterances, samples of which can be found in **Table 4.3** and **Table 4.4**. We also recorded twenty spoken apathetic utterances, samples of which can be found in **Table 4.5**.

The reason for choosing an apathetic control condition over a condition with no speech whatsoever was was to maintain a similar illusion of the robot's intelligence across all three conditions.

The robot's responses for each condition were chosen as follows:

- **Emotionally Appropriate Responses** – The robot spoke with one of the prerecorded responses from the appropriate emotional category, determined by the score the robot received during that demonstration. Among the responses in that category, one was chosen at random without repeating any responses per participant per song. See Table 4.3 for a sample.
- **Often Emotionally Inappropriate Responses** – The robot spoke with one of the prerecorded responses from a random emotional category. Among the responses in that category, one was chosen at random also without repeating any responses per participant per song. See Table 4.4 for a sample.
- **Apathetic Responses** – The robot spoke with one of the prerecorded responses from the apathetic group, again chosen at random without repeating any responses per participant per song. See Table 4.5 for a sample.

In every instance above in which we discuss random choices, it is important to note that across all participants the seed value for the pseudorandom number generator was held constant per dance and per demonstration. The "random" choices above are random in the sense that we did not choose the ordering ourselves, but those choices were constant across all participants, per dance per demonstration.

“Emotionally Appropriate” Condition

Score	Robot’s Response
20	“Oh no, ohh no."
22	“Ugh, man, this is hopeless."
82	“Ooh, check *that* out, we did great!"
89	“Now, how great is that."
91	“Cool, cool, we did well."
94	“Oh yeah, that’s right! Un-huh!"
95	“Oh yeah, oh yeah, oh yeah."
97	“Yeah, well, I’m really good at this."
99	“Cool, cool, we did well."

TABLE 4.3: Sample of the robot’s responses to its scores in the “emotionally appropriate responses” condition. Compare with sample responses in the other two conditions, found in **Table 4.4** and **Table 4.5** below.

4.3 Participants

There were 62 participants, between 18 and 40 years of age, all from New Haven, CT. Most participants were undergraduate and graduate students, none of whom were computer science majors. Our exclusion criteria were lack of English fluency or prior academic experience with robots or artificial intelligence (i.e. students having taken or currently taking a robotics or artificial intelligence course).

“Emotionally Inappropriate” Condition

Score	Robot’s Response
20	“Look at that! That is an awesome score."
22	“Augh, that was bad, that was really bad."
82	“Oh no, that was terrible!"
89	“Oh yeah, that’s right! Un-huh!"
91	“Ugh, I’m so mad!"
94	“Ooh, we’re doing really well."
95	“Hey, that score’s pretty darn good."
97	“Now, how great is that."
99	“Ugh, oh no, I’m so sorry!"

TABLE 4.4: Sample of the robot’s responses to its scores in the “often emotionally inappropriate responses” condition. Compare with sample responses in the other two conditions, found in **Table 4.3** above and **Table 4.5** below.

4.4 Procedure

The participant was told that the purpose of this study was to help the robot learn to dance. The participant was informed of the features of the instruction interface and how to perform the dances. Participants were left alone with the robot and asked to remain standing on the Wii Fit Balance Board Peripheral, positioned in front of the robot, throughout the experiment. Participants would click on buttons on the interface with a mouse that extended to within reach of the Balance Board. After each demonstration,

“Apathetic” Condition

Score	Robot’s Response
20	“We did okay."
22	“Mhmm. That makes sense."
82	“Sure. I’ll take it."
89	“That looks alright to me."
91	“That was... that was okay."
94	“Oh. That’ll do."
95	“Hmm. Looks like we’re doing fine."
97	“That’s decent."
99	“I think that’s fine."

TABLE 4.5: Sample of the robot’s responses to its scores in the “apathetic responses” condition. Compare with sample responses in the other two conditions, found in **Table 4.3** and **Table 4.4** above.

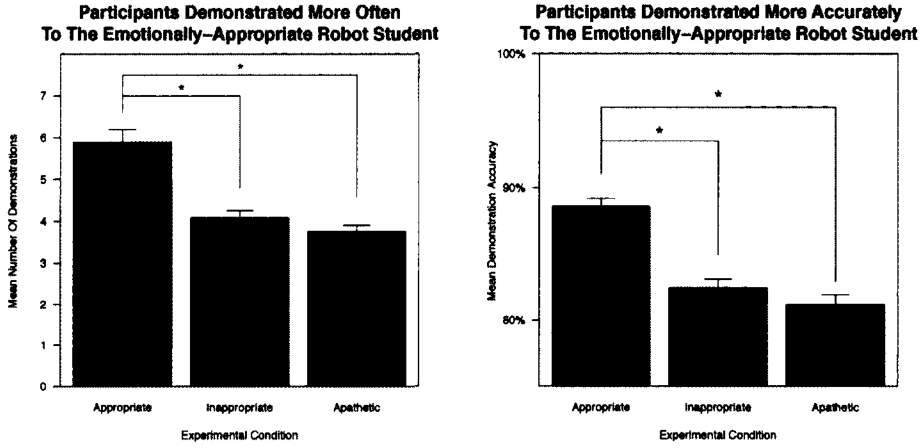
the robot gave its emotional response to the score and, afterward, participants were presented with two buttons, one marked “Move On” and a larger one marked “Teach Again,” as depicted in **Figure** 4.2d.

Participants demonstrated the dance moves in front of the robot as the robot also danced. Participants could choose after each demonstration whether to repeat the same dance or to move on to the next dance, with no option to return to previous dances. Some participants asked the experimenter, during the explanation of instructions, what scores were required or desirable, to which the experimenter consistently replied by requesting that the participant continue his or her demonstrations until he or she felt satisfied with the robot's performance or score. The experimenter did not mention the emotional aspect of the robot's behavior. The experimenter also did not reveal that the participants' accuracy was being scored. The complete text of the instructions is provided below:

"Today you're going to teach our robot, Kate, to dance. We made five thirty-second dances that we want her to learn, each dance is set to its own pop song. Your job is to demonstrate the dances for her and she'll learn them by imitating you. The dances themselves are really simple, they're made up of two kinds of moves: leans and bounces. The screen behind Kate will tell you what moves to do and what moves are coming up. Here, let me show you." At which point a 15 second video was shown demonstrating "leans" and "bounces."

"Each time you dance together, Kate will get a score out of 100%. After each dance, you'll be asked whether you want to teach Kate that song again or to move on to the next song. You can teach each song as many times as you want, the more times you demonstrate it for her the better she'll do. Once you move on from a song, you can't go back to it. Does that make sense? Do you have any questions?" After the participant's questions were answered, they were asked to stand on the Wii Fit Balance Board Peripheral and begin the experiment.

After participating in the study, participants were asked to complete a survey consisting of six open-ended questions followed by two Likert-scale rating questions. The open ended questions were designed to give the impression that the experiment was investigating how well the robot learned the dance moves (e.g. "In your opinion, how well did Kate learn?",



(A) Participants who taught the robot student with emotionally-appropriate responses did significantly more demonstrations that participants in the other two groups, $p < 0.001$.

(B) Participants who taught the robot student with emotionally-appropriate responses performed each demonstration significantly more accurately that participants in the other two groups, $p < 0.001$.

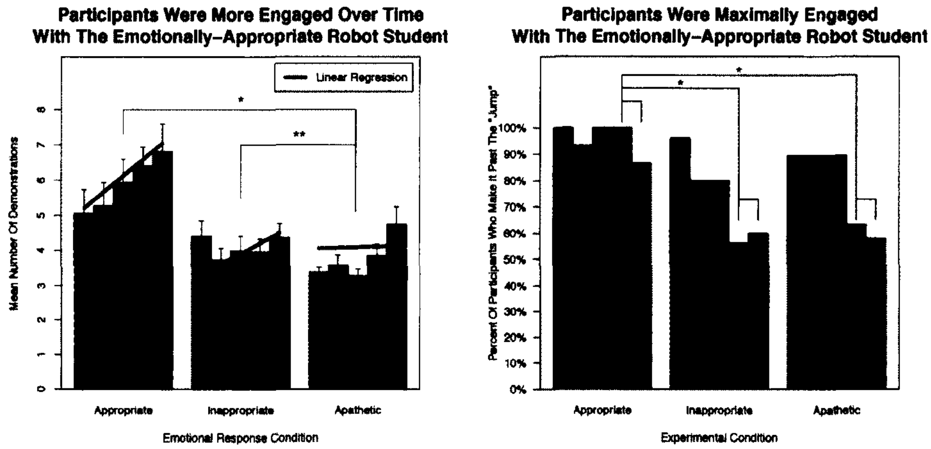
FIGURE 4.3: Our results indicate that participants who taught the robot student with emotionally-appropriate responses performed significantly more demonstrations and performed each demonstration significantly more accurately that participants in the other two groups. (Error bars plot standard error.)

"Do you think you demonstrated the dances well enough?", "What factors influenced your decision to move on from one song to the next?"). The two Likert rating questions were, rating (1) "Kate's emotion responses to her scores...", on a scale of "1 – seemed arbitrary." to "7 – seemed believable.", and (2) "Overall Kate learned...", on a scale of "1 – very poorly." to "7 – very well."

4.5 Results

This study was designed to investigate whether human tutors personalize their tutoring to robots that perform exactly the same in learning tasks but vary in their emotional responses during tutoring. To investigate this, the mean number of demonstrations per dance, over all five dances, was compared across conditions, see **Figure 4.3a**. Participants who taught the robot student with emotionally appropriate responses demonstrated the dances ($M = 5.9, SD = 2.3$) significantly more frequently than those who taught the robot student that gave often emotionally inappropriate responses ($M = 4.1, SD = 1.5$), $t(123) = 5.18, p < 0.001$ as well as demonstrating the dances significantly more frequently than those in the apathetic responses group ($M = 3.8, SD = 1.0$), $t(110) = 6.32, p < 0.001$, and. No significant difference was detected between the apathetic response condition and the often emotionally inappropriate response condition. This result indicates that human tutors do change their behavior based solely on the emotional output of their students.

The mean accuracy of each participant's demonstrations, calculated as described in **Figure** 4.2.2, produced similar results, see **Figure 4.3b**. Participants who taught the robot student with emotionally appropriate responses earned significantly higher accuracy scores ($M = 89\%, SD = 12\%$) than participants in both the apathetic responses group ($M = 81\%, SD = 15\%$), $t(692) = 7.6, p < 0.001$ and the often emotionally inappropriate responses group ($M = 80\%, SD = 15\%$), $t(648) = 6.86, p < 0.001$. Again, no significant difference was found between mean accuracies of participants in the apathetic



(A) The number of demonstrations participants made over time grew fastest in the emotionally-appropriate robot student group. The mean slopes are compared here of linear regressions, single asterisk indicates significance, $p = 0.05$, double asterisk indicates moderate significance, $p = 0.07$.

(B) Participants who taught the robot student with emotionally-appropriate responses were significantly more likely to be patient enough with the robot student to reach the jump in the scores, from below 30% to above 75%. Asterisks indicate significant differences among means, $p \leq 0.01$.

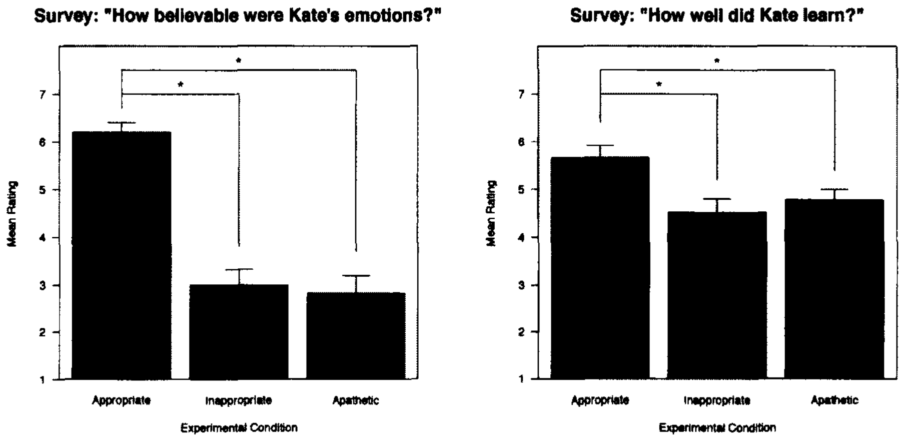
FIGURE 4.4: Results from the behavioral data.

group and the often inappropriate emotional group. This further confirms that human tutors personalize their tutoring based on emotional feedback of students.

For each dance, the robot received only scores that were below 30% until, after some number demonstrations per song, the scores it received would jump to exclusively above 75%. The number of demonstrations necessary to reach that jump in scores was consistent per song across all participants; it was the same for the first three dances and it increased in the fourth and fifth dances. We investigated the percentage of participants that performed enough demonstrations to earn a high score on the last two "increased-difficulty" dances, plotted in **Figure 4.4b**. The percentage of participants in the emotionally appropriate responses group who reached those jumps (93%) was significantly larger than the percentage of participants in the apathetic response group (61%), $t(24) = 2.7, p = 0.01$, and in the often emotionally inappropriate response group (58%), $t(34) = 3.5$, $p = 0.001$. This indicates that participants who taught the robot student with emotionally appropriate responses were not only more engaged with the robot, but more patient with it when it failed.

We also investigated the rate of change of the number of demonstrations over the five dances, between conditions; see **Figure 4.4a**. Fitting each participant's number of demonstrations per dance with a least squares linear regression allowed us to investigate the participant's engagement over time, by comparing the mean slopes between conditions. The mean slope of participants in the emotionally appropriate condition ($M = .46, SD = .70$) was significantly larger than the mean slope of those in the often emotionally inappropriate response group ($M = .02, SD = .59$), $t(26) = 2, p = 0.05$. The mean slope in the apathetic group ($M = 0.30, SD = .41$) was larger than the mean slope in the often emotionally inappropriate group with only moderate significance, $t(41) = 1.8, p = 0.07$. This result tells us that participants who taught the robot student with emotionally appropriate responses were more engaged, more patient, and more consistent than participants in the other two groups.

We also found that even by the end of the first dance, where on average across all groups, participants saw only 4.2 of the robot's responses ($SD = 2.0$), and yet, by the end of the first dance there were already significant differences within the mean



(A) This question verifies our main manipulation: the emotional content of the robot student's responses was correctly identified by participants as appropriate or often inappropriate, in those groups respectively.

(B) The perception of the robot student's successfulness was significantly higher for participants who taught the robot student that gave emotionally appropriate responses.

FIGURE 4.5: Results from the survey data.

number of demonstrations across groups. After the first dance, both appropriate ($M = 5.1, SD = 2.6$) and often inappropriate ($M = 4.4, SD = 2.2$) emotional response groups had a significantly higher number of demonstrations than the apathetic group ($M = 3.4, SD = .70$), $t(16) = 2.5, p = 0.03$ and $t(30) = 2.2, p = 0.04$. This indicates that the personalization human tutors do based on a student's emotional feedback happens early and consistently.

The survey results verified our manipulation – the emotionally appropriate response group rated the robot's emotions ($M = 6.0, SD = .77$) significantly more believable than the apathetic response group ($M = 2.8, SD = .97$), $t(24) = 7.93, p < 0.01$, and the often emotionally inappropriate response group ($M = 3.0, SD = 1.4$), $t(37) = 8.36, p < 0.01$.

(See **Figure 4.5a.**) There was no significant difference between the often-inappropriate emotional group and the apathetic group. This result reveals that participants were noticing the emotional feedback of the student, and not only the scores it received.

The survey results also indicated that the emotionally appropriate response group rated the robot's ability to learn ($M = 5.6, SD = .98$) significantly higher than participants in the apathetic group ($M = 4.8, SD = .97$), $t(29) = 2.62, p = 0.01$, and significantly higher than participants in the often emotionally inappropriate response group ($M = 4.5, SD = 1.4$), $t(37) = 3.02, p < 0.01$; see **Figure 4.5b**. This indicates that participants who taught a robot tutor with emotionally appropriate responses perceived the robot to be smarter than participants in the other two groups perceived the robot student to be, even though the robot performed just as well in all three groups.

4.6 Discussion & Design Guidelines

The central finding of this work is that, even when students perform exactly the same across all learning tasks, naive human tutors significantly alter their tutoring to adapt to students with different emotional responses. Participants taught significantly more often and significantly more accurately to robot students with emotionally appropriate responses. Our first design guideline in this chapter is therefore: if automated personalization systems are to draw the best qualities of human tutoring, they need to pay close attention to the student's emotional feedback. We found that this effect, in which participants treated the emotionally appropriate robot student differently than the others, was robust in several measures including how patiently participants waited for the "jump" in scores, as well as how the number of demonstrations they chose to do grew over time faster than in the other two groups. Though we do not suggest that automated tutors are disengaged with students who behave apathetically or emotionally inappropriately, we do suggest that identifying these behaviors and responding as an expert tutor would, by perhaps questioning the behavior or offering a break, is important to the success of automated tutoring.

Comparing the apathetic condition to the often inappropriate emotion condition, the majority of the statistical analysis supports the null hypothesis – namely, that neither produces significantly different quantity or quality training data. The only exception present is the mean slope data, which produced a marginally significant result between these two groups ($p = 0.07$). (See **Figure 4.4a.**) This trend may indicate that there is some underlying difference that we do not yet have enough statistical power to determine. However, this result also strengthens the design guideline we are proposing: we found that even apathetic responses are enough to cause human tutoring personalization, and that this personalization is similar to the way human tutors treat often emotionally inappropriate robot students. In other words, the emotional output of a student is an incredibly important signal to a tutor and ignoring it by thinking of students as apathetic or unemotional leaves automated tutoring systems without a rich source of data on which to personalize an interaction.

Even though all three of the robot students performed each dance the exact same way and received the exact same scores, the survey data, which can be found in **Figure 4.5b,** indicate that participants in the emotionally appropriate response group believed the robot learned significantly better, on average, than those in either of the other groups, $p < 0.01, p = 0.01$. This result may be due simply to the relative patience of participants in this group, as indicated by their performing more demonstrations and, thus, earning higher scores. Even if that is the underlying cause, this data indicates that human tutors perceive students of the same objective ability as having differing abilities based on their emotional output. This leads to our other design guideline: if automated personalization systems are to be more human-like in their personalization, they should take the emotional signal into consideration when assessing the otherwise objectively-measured skills of a student. This may seem counterintuitive, that skills perhaps should not be objectively measured, but another way to understand this result is that communication skills are sometimes almost as important as the content itself. If a student is not able to establish a good rapport with the automated tutor, because he or she is bored or unenthusiastic and therefore his or her emotional responses are not ideal, the tutor should notice that and report it as an element of the evaluation of that student. Human tutors allow this to affect our judgement of a student but perhaps an automated system can more easily separate the two.

4.7 Conclusion

In this study we investigate how human tutors personalize their instruction to students of the same ability but different emotional responses. Participants were asked to teach the robot several "dances" by demonstrating them repeatedly for the robot student. The robot student then responded to the score it received during each demonstration in one of ways: either with (1) emotionally appropriate responses, (2) often emotionally inappropriate responses, or (3) apathetic responses. We found that participants taught significantly more often and significantly more accurately to the emotionally appropriate robot student than to the other two robot students. We also found that participants who taught the robot student with emotionally appropriate responses rated the robot student as "able to learn" significantly better than participants in either two groups, perhaps because they were less engaged with the student and observed fewer successes, indicating that the emotional responses affect a naive human tutor's perception of a student's ability to learn. We propose design guidelines for future work in automated personalization systems based on these data:

- We suggest that if automated personalization systems are to model the best qualities of human tutoring, they need to pay close attention to the student's emotional feedback, which many systems currently do not do. We found that the emotional output of a student is an incredibly important signal to human tutor, where even apathetic responses caused novice humans to respond differently to students.

- We suggest that if a student is not able to establish a good rapport with the automated tutor, because he or she is bored or unenthusiastic and therefore his or her emotional responses are not appropriate, the tutor should detect this and attempt to intervene.

Chapter 5

The Effect of Personalization in Short-Term Robot Tutoring

In this chapter we investigate to what extent the personalization of a robot tutor can affect student learning gains over the course a single tutoring session. As the first group to investigate the role of personalization in robot tutoring, we were interested in establishing a minimum threshold for the effects of personalization. Is it difficult or expensive to build robot tutoring personalization systems that tailor their output to individual student's strengths and weaknesses? Is it worth creating a personalization system for just a single-session application with a robot tutor?

In this chapter, we show that personalization can be done relatively simply and can make a significant difference even over the course of just one session with a robot tutor. We present two personalization systems we authored for short-term, single-session robot tutoring interactions and compare their effectiveness. We find that both produce significant learning gains: Participants who received personalized lessons from a robot tutor based on these systems performed between 1.0 and 1.4 standard deviations above the mean of participants who received non-personalized lessons from the same robot tutor, corresponding to learning gains in the 84th and 92nd percentile respectively. Participants who received personalized lessons performed between 1.2 and 1.7 standard deviations above the mean of participants who received no lessons whatsoever, corresponding to gains in the 88th to 96th percentiles.

To study the effect of personalization in a single session of robot tutoring, we designed an experiment with four conditions: (1) a condition in which participants received personalized lessons from a robot tutor based on our first personalization system, (2) a condition in which participants received personalized lessons from a robot tutor based on our second personalization system, (3) a condition in which participants received non-personalized lessons from the same robot tutor as in the first two conditions, and (4) a condition in which participants were asked to perform the same learning tasks as in the previous three conditions but with no lessons or tutoring whatsoever. We find that personalization has a significant impact on student learning outcomes in robot tutoring, even in the course of just one session with the robot. In our work this impact is an average of 1.2 standard deviations over the mean performance of participants who received non-personalized instruction and 1.5 standard deviations over the mean performance of participants who received no lessons whatsoever. We compare the two personalization systems below and provide guidelines for future work in short-term personalization in robot tutoring.

5.1 Introduction

We are the first to study personalization in the context of robot tutoring, whereas previous work has explored personalization in other kinds of human-robot interactions. The most significant project in this area is Snackbot, a robot that personalizes dialogue in reference to an individual user's history of snack choices (i.e. an apple versus a candy bar) (Lee et al. 2012). When Snackbot personalized its interactions it was found to be more engaging by participants than a non-personalized version of the robot, leading to an increased desire to use it and an increase in social behavior directed towards the robot.

In other work that features personalization, Kidd and Breazeal (2008) present a robot weight loss coach that generates customized dialogue based on the self-reported progress of the user, finding that the physical embodiment of a robot coach produces significantly more engagement with the robot. This project does not specifically isolate the role of dialogue personalization. Leite et al. (2012) conducted a long-term study of elementary students playing chess with help from a robot, exploring how supportive the students perceive the robot tutor to be depending on the kind of feedback it gave students. The robot chess tutor did not personalize the kind of support it gave but this study asked students what kinds of support they preferred, which future work could use to personalize robot tutoring interactions. In the work of Sung, Grinter and Christensen (2009), users that decorated, and thus "personalized," their Roombas, self-reported higher engagement with the robot and more willingness to use the robot in the future. These studies indicate that personalization in human-robot interaction produces better user engagement across many kinds of interactions.

We are the first to study whether this increased engagement produced by personalization leads to learning gains for students interacting with a robot tutor. The most significant earlier robot tutoring project is called RUBI, a robot tutor intended for early childhood education (Movellan et al. 2007). The RUBI project spans a variety of educational and robotics research but the RUBI group has not investigated the role of personalization. We provide an overview of their contributions in Chapter 1, Section 1.4.

We present two systems in this chapter that are intended to evaluate the effectiveness of personalization in robot tutoring over a single session. We find that personalization in robot tutoring can be effective even with relatively simple systems and in just one hour-long tutoring interaction.

5.2 Overview

This study investigates the effect of personalization on student learning outcomes in a single session of robot tutoring. The curriculum the tutor teaches and the apparatus of this study are the same as in our earlier work investigating the effect of embodiment in robot tutoring in Chapter 2. A description of the curriculum and apparatus can be found in Section 2.2 and Section 2.3.

We provide a summary of key details here. Each participant in this study was asked to solve the same series of four logic puzzles called 'Nonograms' (also called 'Nonogram puzzles'). Nonograms are Japanese fill-in-the-blanks grid-based puzzle games that require players to perform many layers of logical inference to complete, similar to Sudoku. A sample Nonogram puzzle, with solution, can be found in **Figure 2.2**.

There were four conditions in this study: two in which participants received personalized Nonograms puzzle-solving lessons, one in which participants received non-personalized Nonograms puzzle-solving lessons, and the last condition in which participants solved the same series of four Nonogram puzzles with no assistance. In the three conditions in which participants received robot tutoring, periodically as participants solved the puzzles, the robot tutor interrupted them to deliver a Nonograms puzzle-solving lesson. These puzzle-solving lessons consisted of pre-recorded audio with synchronized lesson-specific on-screen visual aids, each lasting between 21 – 47 seconds, and each describing a unique Nonograms puzzle-solving skill. A transcription of the audio content of these lesson can be found in **Section 2.3**.

The only difference between the three conditions that received robot tutoring was the ordering of the lessons. In the personalized conditions, the lessons were ordered based on one of the two personalization systems we created. In the non-personalized condition, the lessons were chosen randomly among the lessons that applied to the current state of the game board when each lesson was given. We measured how quickly participants were able to solve each of the four puzzles; the faster a participant was able to solve puzzles, the better at the puzzle-solving skills we assessed them to be. We compared the mean puzzle-solving times between participants across all four groups to evaluate the effect of our two personalization systems on short-term robot tutoring.

5.3 Curriculum: 'Nonograms'

This study uses the same curricular domain, Nonograms, as we used in **Chapter 2**. We also use the same ten Nonograms puzzle-solving skills and ten pre-recorded lessons we authored for that earlier work. See **Section 2.3** for a description of the rules of Nonograms and the skills and lessons we created.

In this study we maintain the manipulation from our earlier work in which the fourth puzzle was a disguised 90° rotation of the first puzzle. Most Nonogram puzzles requires a slightly different subset of skills to solve, applied in a different order between puzzles. Because the first and fourth puzzles in this study had functionally identical gameboards, they required exactly the same subset of Nonograms puzzle-solving skills, in the exact same order, to complete successfully. As a result, comparing the performance of the a single participant on the first puzzle to that same participant's performance on the fourth puzzle allows us to evaluate that participant's skill competency growth over the course of the study. We use this within-subjects measure to evaluate the effectiveness of our personalization systems on student learning outcomes below.

5.3.1 Skills & Lessons

For our work in **Chapter 2**, we identified ten unique Nonograms puzzle-solving skills and we recorded ten lessons, one for each skill, each under a minute in length. Each of these skills and corresponding lessons are described in **Section 2.4**. We describe how these lessons were ordered, depending on experimental condition, below.

It is significant to note that these skills are not trivial to order because, as is the nature of tutoring any cognitively-challenging task, it is difficult to know what skill a student has full knowledge of, and what skill a student may be lacking in. One can observe how the student solves problems, but mapping the observations one makes to the skill competency of an individual student is a difficult task because there is a many-to-many relationship between such observations and such skill. Creating this mapping, and inferring from it the skill competency of an individual student, is a necessary job for personalizing the ordering of the instructional content. Below, we present two initial efforts to accomplish this goal.

5.4 Conditions

There were four conditions in this study, two conditions in which participants received personalized lessons from a robot tutor, one condition in which participants received non-personalized lessons from a robot tutor, and a condition in which participants solved the puzzles with no tutoring at all. Each participant experienced exactly one of these conditions.

5.4.1 No Lessons

To assess a baseline of performance on this puzzle-solving task, we measured how well participants learned Nonograms puzzle-solving strategies on their own, with no instruction. Collecting this data allows us to compare the effect of our personalization systems against participants' own efforts at inferring the puzzle-solving skills over the course of the session, which allows us to quantify the impact of our personalization systems. We expected that participants would get better at the task over time, inferring the same skill-solving strategies but more slowly and perhaps less clearly. In this condition, there was no robot present, so the results represent a baseline measure of how participants solve this task in a non-social setting, simply thinking the puzzles through on their own.

5.4.2 Non-Personalized Lessons

To isolate the effect of personalization in a single session of robot tutoring, we designed a condition in which participants received the same pre-recorded lessons as in the personalized conditions but ordered randomly. When the tutor gave a lesson in this condition, it picked randomly from among the subset of lessons that could be directly applied to the current gameboard state at the time the lesson was given.

An alternative strategy for the non-personalized condition would have been to order the lessons in a fixed, pre-defined sequence. Doing this was not feasible here because Nonograms skills can be applied in slightly different orders but result in the same solution. A fixed curriculum in this domain would have delivered lessons that were applicable to some participants' gameboards when they were delivered but not to others, rendering a comparison between non-personalized and personalized lessons as a comparison of irrelevant and relevant lessons. Instead, with our system, we can isolate the personalization while keeping the relevance of the lessons constant across all conditions.

5.4.3 Personalized Lessons

We personalized the lesson ordering based on two models that estimate the skill competence of each participant on each of the ten skills. The job of the model is to interpret what each move a participant makes indicates about the underlying internal cognitive processes of that participant. A significant challenge in modeling these cognitive processes is that each of the participant's actions can reflect the presence or absence of many skills at once. We describe two algorithms that attempt to unscramble these potentially mixed signals here.

It is important to note that the lesson ordering algorithms presented here are not proposed as optimal solutions to the lesson ordering problem more broadly. Instead, we are interested specifically in isolating what effect, if any, relatively simple personalizations can have on a single sessions with a robot tutor. Both algorithms take as input the moves participants make in the puzzles and produce as output a single skill in which the model indicates a participant's knowledge is lacking.

5.4.3.1 Additive Model

In this model, each individual Nonograms puzzle-solving skill i has an associated function S_i that defines precisely what it means to apply skill i. Each function S_i takes as input a potential state of the gameboard or world ($w_t \in W$) and returns a set, $\{w_{t+1}, w'_{t+1}, w''_{t+1}, w'''_{t+1}, ...\}$ that contains all of the possible resulting world states after skill i applied is applied to world state w_t.

$$S_i(w_t) = \{w_{t+1}, w'_{t+1}, w''_{t+1}, w'''_{t+1}, ... \mid \text{skill } i \text{ was applied to world state } w_t\}$$

We say skill i is "not applicable" to a world state w if and only if $s_i(w_t) = \{w_{t+1} \mid w_t = w_{t+1}\}$. In other words, if applying skill i to world state w_t only produces one possible outcome, a state identical to w_t, then skill i does not apply to gameboard w_t. This happens in Nonograms when a skill cannot be used to make progress in any of the rows or columns of a gameboard.

The skill functions S_i are used in two ways, to detect **successful demonstrations** and **missed opportunities**:

- We say skill i was **successfully demonstrated** at world state w_t if $w_t \in S_i(w_{t-1})$. In Nonograms, this represents an instance where the participant performs an action that matches the definition of one of the ten skills we defined.
- We say the participant **missed an opportunity** to demonstrate skill i at world state w_t if the participant takes no action *and* $s_i(w_t) \neq \{w_{t+1} \mid w_t = w_{t+1}\}$.

This occurs when the skill *can* be used to make some progress in a row or column somewhere on the gameboard, but the participant does not make any move.

We incorporate these two applications of S_i as follows. We define p_i as a boolean value that is 1 if and only if skill i has been **successfully demonstrated** in the previous timestep, and n_i as a boolean value that is 1 if and only if the participant **missed an opportunity** to employ skill i in the current timestep. The timesteps for these two booleans are different. The boolean p_i is evaluated every time a box on the Nonograms board is shaded. Therefore, every time a participant makes a move in Nonograms, they have the potential to **successfully demonstrate** one or more of our ten skills. The boolean n_i is evaluated every time the state of the world does not change for a 3 seconds period, and then it is evaluated repeatedly every 1 second thereafter until another move is made. These time delays were chosen based on the authors' subjective experience with the task domain that indicates that a pause while playing Nonograms indicates that the user is stuck, typically after 3 seconds of inactivity, and continuing thereafter, which we sample every second. In practice, this means that after three seconds of inactivity, the model starts to accrue evidence that participants are **missing opportunities** to demonstrate any skills that are currently applicable to the board during their inactivity. These inputs are summed with the following equation, either when a skill is demonstrated or an opportunity is missed:

$$a_{s,t} = \omega_0 + \sum_{i=0}^{t} (\omega_p p_i - \omega_n n_i)$$

Each skill has its own assessment $a_{s,t}$. The values of $a_{s,t}$ vary from 0 to 100. We used three weights in these calculation, ω_0 which is an initial seed value, set to 50 for all skills, and ω_p and ω_n which represented the relative frequency with which we expect to see **successful demonstrations** and **missed opportunities** respectively, set to $\omega_p = 50$ and $\omega_n = 1$. A floor of 0 and a ceiling of 100 was applied to the summed value at each timestep. The weights and seed value used in this algorithm were subjectively derived and fine-tuned based on pilot studies. These pilot studies revealed that participants sometimes took pauses of up to a minute to plan their moves, and as a result, any of the skills applicable to the board in those sixty seconds would decrease by as much as 57. We tuned our weights in this additive model to reflect the notion that a single **successful demonstration** would cancel the effect of a little less than a minute of **missed opportunities**.

The additive model updates each of these skill assessment scores, $a_{i,t}$, as participants solve puzzles. When choosing a lesson based on this model, we choose the lesson associated with the skill with the lowest score. In the event of a tie, we choose randomly among the lessons associated with the tied lowest scoring skills.

5.4.3.2 Bayesian Model

A weakness of the additive skill assessment algorithm is its susceptibility to local maxima and minima. When individual skill assessments reach floor or ceiling, the additive algorithm essentially ignores the participants' performance history. A good human tutor does not forget previous successes or failures in light of more recent observations.

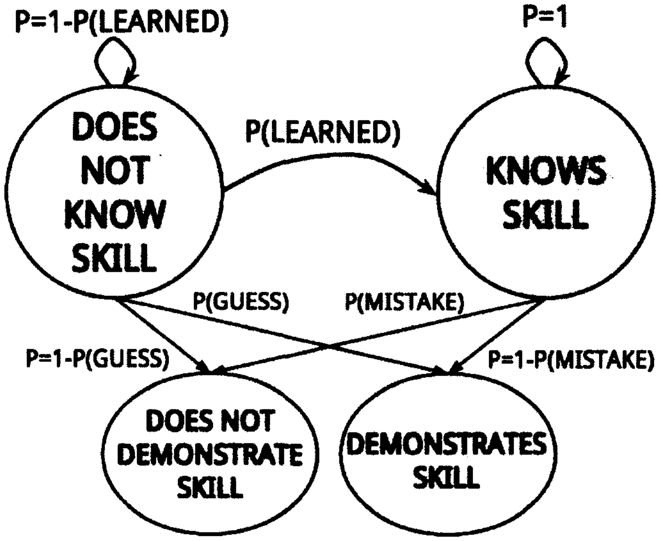


FIGURE 5.1: The Hidden Markov Model used for each skill in the Bayesian personalized tutoring condition. P(LEARNED) is the likelihood that participants learned the skill at a given timestep, P(MISTAKE) is the likelihood that a participant who knows a skill makes a mistake and does not apply it, and P(GUESS) is the likelihood a participant does not know a skill but guesses the right answer. These parameters were learned per participant, per skill. More details can be found in **Section** 5.4.3.2.

We addressed this weakness by offering a Bayesian network approach, in the form of Hidden Markov Models (HMMs). We created one HMM for each skill for each participant, in the form illustrated in **Figure** 5.1. These HMM had two hidden states: either the participant (1) knew the skill or (2) did not know the skill. There were two possible observations, either (1) participants demonstrated a skill, which is defined in the same way as the **successful demonstrations** are defined above, or (2) participants did not demonstrate a skill, defined the same way as the **missed opportunities** above.

Because the exclusion criteria for this study included any previous experience with Nonograms, we knew the initial distributions of the hidden states for all ten skills: we set the probability that any participant knew any of the ten skills at the beginning of the study to 0%. For each participant and each skill, there were only three parameters to learn, P(LEARNED), P(MISTAKE), and P(GUESS) defined in **Figure 5.1**. These were learned and updated at every timestep with the well-known Baum-Welch algorithm, an overview of which can be found in Welch (2003).

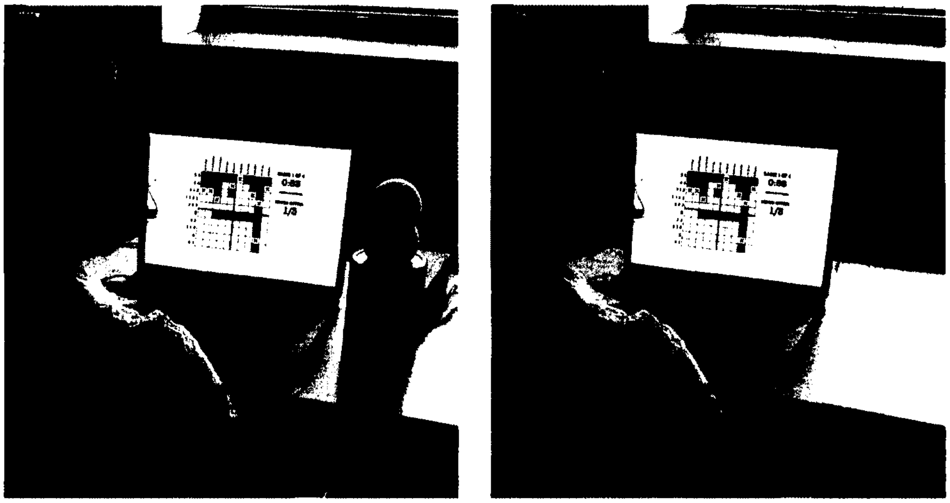
The parameter P(GUESS) serves two purposes in this work. First, though we actively discouraged participants from guessing throughout our instructional materials, some participants who were stuck on a puzzle for a long time did guess. We discouraged guessing primarily to avoid the situation in which an incorrect guess renders the puzzle unsolvable until that move is undone. However, if a participant does guess incorrectly, the subsequent moves the participant makes are still modeled correctly. The skill functions S_i, defined above, upon which the observational states in the HMMs are based, only depend on the state of any one row or column at a time. A guess that incorrectly shades a box in any given row or column still allows subsequent moves to be modeled by the same skill functions S_i.

In addition to modeling guessing, P(GUESS) also allows us to model when a given observation could be interpreted as evidence of more than one skill. In Nonograms, some moves a participant makes cannot be identified as a demonstration of one skill, but rather as one of a set of skills. In these situations, the HMMs for all the potential skills are given the input that the participant demonstrated that skill, even though it is not clear from the move he or she made whether a participant knows one of those skills, some subset of those skills, or all of those skills. In this sense, $P(\text{GUESS})$ is the likelihood that a participant guesses or demonstrates a related skill. What these two events have in common is that we are not sure whether the participant knows the skill or does not know the skill, though we have some evidence that the skill was demonstrated. This is why we chose HMMs to model this phenomena, and the Baum-Welch algorithm to learn the transition probabilities over time.

The output of these ten HMMs was calculated with the Viterbi algorithm, which finds the most likely sequence of states that explain a given sequence of observations (**Forney Jr. 1973**). When the robot tutor gave a lesson using this model, it chose randomly among the skills for which the Viterbi algorithm predicted that the participant was in the "does not know skill" hidden state.

5.5 Robot

We used the same robot to deliver these lessons as in **Chapter 2**. **Section 2.5.1** describes the robot and its behavior in this context. **Figure 5.2** shows the two apparatuses of the experiment, with the robot tutor placed beside the full-screen Nonograms graphical user interface in three of the four conditions, and solely the full-screen Nonograms program with no robot tutor in the fourth condition.



(A) Apparatus in three of the four conditions: two conditions in which the robot provided personalized lessons to participants and one in which the robot provided non-personalized lessons.

(B) Apparatus in one of the four conditions: participants solved the same series of four 'Nonograms' puzzles with no tutoring or assistance whatsoever.

FIGURE 5.2: Apparatuses of the four conditions in this study. Three of the four conditions involved a robot tutor, as in **Figure 5.2a**, one did not, as in **Figure 5.2b**.

5.6 Participants

There were 80 participants in this study, 20 per condition, all of whom were between 18 and 42 years of age. Most participants were undergraduate and graduate students of Yale University. Exclusion criteria for participants were lack of English fluency, prior academic experience with robotics or artificial intelligence, and prior experience with Nonograms.

5.6.1 Procedure

Before participating in this study, participants read a two page instruction manual teaching them the rules of Nonograms and watched a two minute instructional video teaching them how to use the computer interface that we designed. In these instructional materials, participants were encouraged to use logical reasoning to make moves in the game, rather than guessing. Afterwards, any questions about the puzzle game and experiment were answered by an experimenter. The text of the instructional manual can be found in Section 2.8.

During the experiment, participants were alone in a room with the robot, the computer, and a video camera positioned behind them, see Figure 5.2. Participants chose when they were ready to start each new puzzle. Games ended either when the participant solved the puzzle or when fifteen minutes had elapsed, whichever came first.

After the conclusion of the final puzzle, participants were asked to complete a survey consisting of five Likert-scale questions with open-ended follow-up questions for each. The questions were designed to assess whether the lessons were helpful, clear, and influential, as well as the user's perceptions of the tutor. We asked participants to rate: how relevant the lessons were, how much the lessons influenced their gameplay, how well participants understood the lessons, and how "smart/intelligent" and "distracting/annoying" they perceived the tutor to be. The intention of these questions was to reveal differences between tutoring conditions that could explain any performance differences between groups.

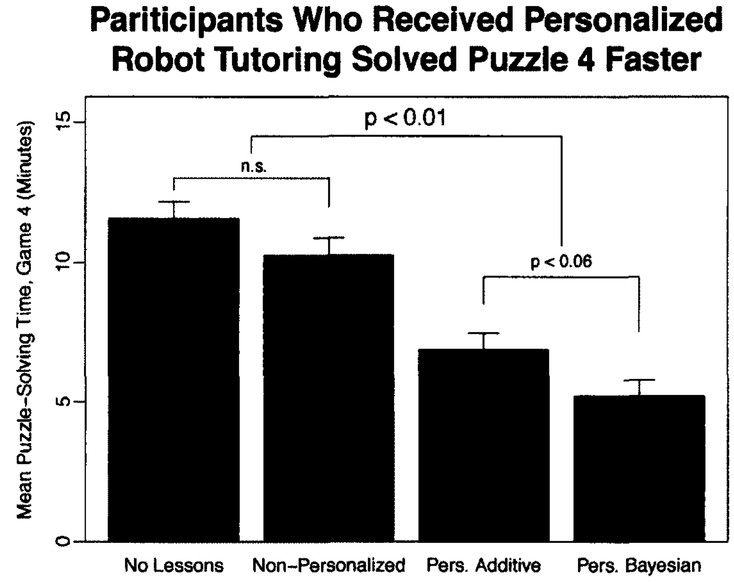


FIGURE 5.3: Participants who received personalized robot tutoring solved the fourth puzzle significantly faster, $p < 0.01$, on average than participants in the two control groups, on average. Participants who received personalized tutoring based on the Bayesian model outperformed participants who received tutoring based on the Additive model with borderline significance, $p < 0.6$. (Error bars in graph plot standard error.)

5.7 Results

This study investigates whether personalization of robot tutors can make an impact on student learning outcomes over the course of a single session. The task-performance measure we used to evaluate the impact of the tutoring is the length of time participants took to solve each of the four puzzles. For the purposes of calculating means, puzzles that were not yet completed at the fifteen minute time limit were scored as having been completed in fifteen minutes. The rate of failure was not significantly different between groups for any of the four puzzles, across the four conditions, varying from 29% to 38% in the first puzzle to 9% to 17% in the fourth.

Participants who received personalized lessons, when taken together, solved three of the four puzzles significantly faster, on average, than those who received either non-personalized lessons or no lessons at all, taken together: $t(38) < 0.03$ for the second game, $t(39) < 0.01$ for the third game, and $t(39) < 0.001$ for the fourth. The means and standard deviations can be found in **Table 5.1**, plots can be found in **Figure 5.3** above and **Figure 5.4** below. These results indicate that our personalization systems produced significantly improved student learning outcomes over non-personalized tutoring, in just a single session with the robot.

	Puzzle 1	*Puzzle 2*	*Puzzle 3*	*Puzzle 4*
No Lessons	13.6 ± 2.2	13.0 ± 2.3	12.3 ± 2.5	11.6 ± 2.7
Non-Personalized Lessons	13.8 ± 1.4	12.5 ± 2.0	11.4 ± 2.3	10.3 ± 2.9
Personalized Lessons, Additive	12.7 ± 2.6	10.0 ± 3.5	9.4 ± 3.0	7.6 ± 3.1
Personalized Lessons, Bayesian	12.2 ± 2.3	9.8 ± 2.4	6.9 ± 3.4	5.2 ± 2.6

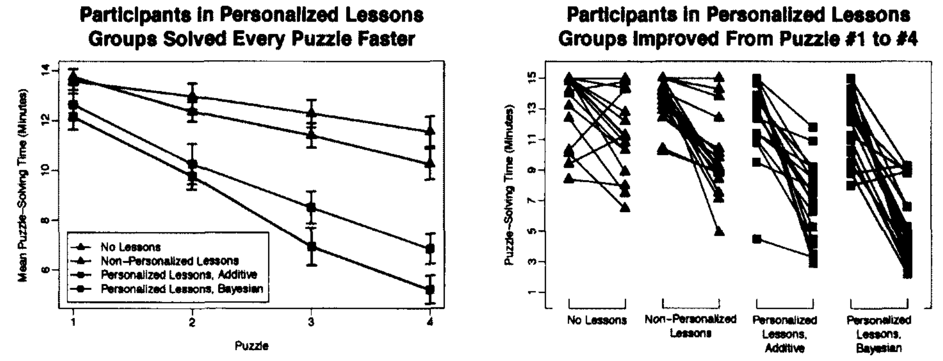
TABLE 5.1: Puzzle-solving times given in means and standard deviations, measured in minutes. In each puzzle except the first, participants in both personalized lessons groups solved the puzzle significantly faster than participants in both the non-personalized lessons group and the no lessons groups. **Table 5.2** below compares performance in the fourth puzzle across conditions.

	Personalized Lessons, Additive	*Personalized Lessons, Bayesian*
Non-Personalized Lessons	Personalized group improved by $Z = 1.0$, 84^{th} percentile.	Personalized group improved by $Z = 1.4$, 92^{nd} percentile.
No Lessons	Personalized group improved by $Z = 1.2$, 88^{th} percentile.	Personalized group improved by $Z = 1.7$, 96^{th} percentile.

TABLE 5.2: Using distribution data in **Table 5.1** above, we present the relative improvements of participants in the personalized conditions compared to participants in the control conditions **on the fourth puzzle**. Presented as Z-scores, corresponding to the number of standard deviations away from the control condition's mean the personalized condition's mean appears in the control condition's distribution. In other words, the percentile tells us where the mean of the personalized condition lands in the control condition's distribution.

Between the two personalized lessons groups, one using the Additive model, the other using the Bayesian model, the group that received personalized lessons based on the Bayesian model did better on the last puzzle ($M = 5.2, SD = 2.6$) than the group that received personalized lessons based on the Additive model ($M = 7.6, SD = 3.1$), with borderline significance $t(37) < 0.05$. See **Figure 5.3** above. For this reason we recommend the Bayesian method for future groups pursuing single-session robot tutoring personalization.

In this study, the fourth puzzle consisted of the same gameboard as the first, disguised

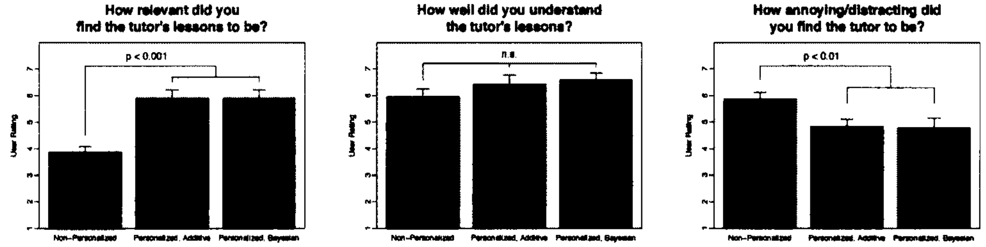


(A) Mean solving time per condition, per puzzle. Participants who received personalized lessons solved each puzzle faster than participants in the non-personalized group, the last three puzzles, significantly faster. See Table 5.1 for data.

(B) The first and fourth puzzles were the same gameboard disguised by a 90° rotation, therefore measuring a participant's learning on the same puzzle-solving skills at the beginning and end of the study.

FIGURE 5.4: Personalization produces greater learning gains, even in one session with a robot tutor: (a) Participants whose lessons were personalized solved the last three puzzles significantly faster than participants in either control group. (b) Participants receiving personalized lessons significantly improved their same-puzzle solving time over participants in either control group.

by a 90° rotation. Different Nonogram puzzles require slightly different subsets of Nonograms puzzle-solving skills to complete and those skills are typically applied in differing orders depending on the puzzle gameboard. Therefore, the difference in completion times between the first and fourth puzzles in this study, in which the gameboards required the exact same subset of Nonograms puzzle-solving skills in the same order, is a within-subjects measure of an individual participant's improvement over the course of the experiment. According to this metric, participants in both personalized lessons groups, when taken together, improved ($M = 5.8, SD = 3.3$) their same-puzzle solving



(A) Participants who received personalized lessons rated the lessons as significantly more relevant than participants who received non-personalized lessons, $t(33) < 0.001$. This indicates that participants felt both personalization systems produced relevant lessons to their needs, as was intended.

(B) Participants who received non-personalized lessons rated their understanding of the lessons highly, and not significantly differently from participants in either personalized lesson condition, despite their gameplay performance indicating that their understanding of the lessons was not as high as the personalized groups.

(C) Participants who received non-personalized lessons did, however, rate rate the robot tutor as significantly more "annoying/distracting" than participants in either of the personalized groups, $t(38) < 0.01$. Perhaps this reflects the lack of relevancy of the lessons in the non-personalized group, leading to participant "annoyance/distraction."

FIGURE 5.5: Survey results comparing non-personalized lessons and the two personalized lessons conditions.

time significantly more than participants in either of the control groups, taken together ($M = 3.1, SD = 2.4$), $t(31) < 0.01$. See **Figure 5.4b**. This is another validation of the effectiveness of our personalization systems.

Survey results indicate that participants in the personalized lessons groups rated the lessons significantly more relevant to them ($M = 4.9, SD = 1.4$) than participants in the non-personalized group ($M = 2.9, SD = 1.1$), $t(33) < 0.001$, as seen in **Figure 5.5a**, which indicates that participants were able to tell when lessons were targeted towards their individual needs. However, there was no significant difference in how participants rated their understanding of the lessons between the personalized groups ($M = 5.4, SD = 1.5$) and the non-personalized group ($M = 5.0, SD = 1.4$), see **Figure 5.5b**. Nor was there a significant difference in how participants self-assessed the degree to which their gameplay was affected by the lessons, between the personalized groups ($M = 4.3, SD = 1.3$) and the non-personalized group ($M = 4.1, SD = 1.3$). These results indicate that participants were able to identify the targeting of the lessons, but not the extent to which the lessons they received impacted their learning.

Participants who received personalized lessons rated the robot as significantly "smarter" or more "intelligent" ($M = 4.7, SD = 1.8$) than participants who received non-personalized lessons ($M = 3.5, SD = 1.6$), $t(36) < 0.03$. Those participants also rated the robot tutor as significantly less "annoying/distracting" ($M = 3.8, SD = 1.2$) than participants who received non-personalized lessons ($M = 4.9, SD = 1.2$), $t(38) < 0.01$, see **Figure 5.5c**. These data indicate that although participants were not able to identify the extent to which the personalization influenced their learning, they did ascribe more positive ("smart") and less negative ("annoying") social characteristics to the robot tutor that personalized its lessons more than to the robot that did not perform personalization.

5.8 Discussion

This study assesses whether relatively simple personalization in robot tutoring affects students' learning outcomes over the course of a single session with the robot. The data indicate that even simple personalization, experienced by participants for only one tutoring session over the course of an hour, can raise mean learning gains by as much as 1.4 standard deviations compared to a non-personalized tutor, see **Table 5.1** and **Table 5.2** above.

An effect size of 1.4 standard deviations, or 1.4 sigma, is more than the mean standard deviation effect size of 0.76 sigma reported by Intelligent Tutoring Systems (ITS's) evaluations, when comparing ITS's to traditional classroom instruction (**VanLehn 2011**). This difference, in part, can be accounted for by the effects of the physical embodiment of the robot tutor. In **Chapter 2** we found this effect to raise learning gains by 0.3 standard deviations in this Nonograms domain over the learning gains made by participants who received an on-screen tutor.

Another potential reason for the size of the effect is the nature of Nonograms, in which a participant's success hinges on several layers of logical inference. It could have been that participants who received personalized lessons caught on to the form of a general Nonograms strategy more quickly than those in the control groups. An early lead in Nonograms puzzle-solving strategies may have allowed these participants to progress faster and perhaps feel more motivated, causing them to widen the performance gap over time between themselves and participants who received non-personalized lessons or no lessons.

The self-report survey data indicate that participants did not report more difficulty understanding the lessons presented to them in the non-personalized condition than in either personalized condition. All three groups rated their own understanding fairly highly: a mean of 5.4 across the personalized lessons groups and 5.0 in the non-personalized group, out of 7, $t(36) = 0.32$. A plot of this data is found in **Figure 5.5b**. It is notable that the non-personalized lessons group reported a relatively high understanding of the lessons despite performing significantly worse than the personalized groups. This may indicate that the population we worked with was reluctant to admit that they did not understand the lessons, in the context of a study. Alternatively, perhaps the participants who received non-personalized lessons did understand the lessons in some sense, but failed to see opportunities in which to apply them.

5.9 Conclusion

In this chapter we investigate the role of relatively simple personalization algorithms in single-session robot tutoring. We compare participants' puzzle solving times across four conditions: two in which participants received personalized lessons from a robot tutor, one in which participants received non-personalized lessons from the same robot tutor, and a condition in which participants solved the same series of puzzles as in the other conditions but with no robot tutor or instructional assistance whatsoever. We find that participants who received even relatively simple-to-achieve personalized lessons for just a single hour-long session significantly outperformed participants who received non-personalized lessons by 1.3 standard deviations on average. We present these results as evidence that personalization can benefit short-term robot tutoring interactions, and that arriving at an effective personalization algorithm may not be as difficult as previously thought.

Chapter 6

The Effect of Personalization in Longer-Term Robot Tutoring

In the previous chapter we designed systems for shorter-term personalized robot-tutoring interactions, those limited to a single session. In this chapter, we describe a system intended for longer-term personalizations, those consisting of more than one but fewer than ten sessions. We present our personalization system which orders curriculum based on an adaptive Hidden Markov Model (HMM) that evaluates students's skill proficiencies and we present a study investigating the effectiveness of this personalization system in a five-session interaction with a robot tutor, taking place over the course of two weeks.

In this work, we challenged ourselves to create an automated robot tutor that could be used in real-world learning task, rather than in contrived laboratory learning task as in our previous chapter. The domain we chose was English as a Second Language

(ESL) education and the population we worked with were native Spanish-speaking 4- to 7-year-olds. We authored an interactive adventure story in Spanish with 24 interchangeable chapters, each offering students a chance to practice one of four English grammar skills. We ordered these interchangeable chapters in one of two ways based on the conditions in our study. Participants either received lessons: (1) ordered by our adaptive HMM personalization system which selects a chapter based on a skill that the individual participant needs more practice with ("personalized condition"), or (2) ordered randomly from among the chapters the participant had not yet seen ("non-personalized condition"). We found that participants who received personalized lessons from the robot tutor outperformed participants who received non-personalized lessons on a post-test by 2.0 standard deviations on average, corresponding to a mean learning gain in the 98th percentile.

6.1 Background

According to the 2010 United States Census data, twenty percent of American households speak a language other than English in the home (**U.S. Census Bureau 2011**). Children raised in non-native English-speaking households face a severe preparatory disadvantage in school relative to their native-speaking peers (**Saunders 1988**). Language-based disadvantages accumulate throughout a student's career and worsen in later grades as reading comprehension becomes more critical to academic success in all subjects (**Callahan 2005**).

The largest population affected by this systemic disadvantage in the United States is Hispanic Americans. Sixteen percent of American households speak Spanish as the primary language in the home (U.S. Census Bureau 2011). The number of native Spanish speakers in the United States has grown by 24 million between 2000 and 2010 (U.S. Census Bureau 2011). Hispanic Americans have the lowest rates of high school and college degree attainment of any racial-ethnic group in America. With less education, Hispanics are at a competitive disadvantage in the workforce. According to the US Bureau of Labor Statistics, Hispanic American unemployment has been roughly 20 to 50% higher than Non-Hispanic American unemployment every year since the data was first collected in 1974 (U.S. Bureau of Labor Statistics 2014).

Effective 'English as a Second Language' (ESL) education is vital to leveling the playing field for children raised in non-native English speaking homes. Though there are many successful programs supplying ESL education across the country, especially in major metro areas like New York and Los Angeles, millions of Hispanic students still receive little or poor-quality ESL education (Humes, Jones and Ramirez 2011).

We envision an in-home robot tutor that can serve as an English-fluent interaction partner for non-native speakers. As a first step towards this vision, we created a robot tutor that provided personalized one-on-one ESL instruction to Spanish-dominant first grade students in a bilingual elementary school.

6.2 Related Work

Younger students, such as those who participated in our study, are still learning their dominant language as well as learning English. For this population there is a targeted research field related to ESL education called 'English Language Learning,' or ELL. ELL programs are similar to ESL programs with the exception that ESL assumes fluency in the student's dominant language, whereas ELL curricula are designed for students who are learning more than one language at a time (Nero 2005). In our discussion of this project, we generalize our results from an ELL population of first grade students to the broader ESL community. We do this because the main measure in this work is correctness of translation tasks from a student's dominant language to English, which is a core competency in ESL research (Auerbach 1993). Our vision for this work is that it will serve both the ELL and ESL populations.

In developing the algorithms necessary for a longer-term personalized automated tutoring interaction, we base our work on that of the automated tutoring systems developed by the Intelligent Tutoring Systems (ITS) community. An overview of these systems can be found in Chapter 1, Section 1.3. For this work, we made a curriculum-sequencing tutor that does not provide step-by-step feedback such as the robot tutoring system we created earlier in Chapter 5. Instead, this tutor sequences an individualized path through available curriculum to maximize the effectiveness of the lessons for each student. For more information about curriculum-sequencing automated tutors, see Section 1.3.1 of Chapter 1.

In addition to the related ITS research, our work is also similar to a body of education research called 'Computer Assisted Language Learning' (CALL), for an interview see Levy (1997). CALL is a branch of education research that studies the effectiveness and implementation of computer-based tools that are intended to assist language learners or teachers, including static resources like webpages and translation software (Levy and Stockwell 2013). A common paradigm in CALL research are systems that process the speech of the user and correct errors in pronunciation, prosody, or grammar (Eskenazi 2009). CALL systems typically do not vary their outputs based on a model of the user, like our automated personalization system does for this robot tutoring intervention. We evaluate the effectiveness of our robot language tutor intervention with a standard pre-test/post-test metric, a common practice in CALL and education research more broadly (Littleton and Light 1999).

In this study, we focus on teaching English as a Second Language to children ages 4 to 7. When learning a second language, age is also very important. The age of first consistent exposure to a second language is the best known predictor of future fluency (Johnson and Newport 1989). This finding influences our choice of target populations for this work, as it indicates that the best time to start teaching a second language is well before puberty, ideally under 9 or 10 years of age (Johnson and Newport 1989). We chose to work with first grade students for this reason.

6.3 Methodology

In this chapter we present the implementation details of our automated personalization system and an experiment in which we evaluate the system's effectiveness in a language learning task with children ages 4 to 7. We authored an interactive adventure story in Spanish with 24 interchangeable chapters, each offering students a chance to practice one of four English grammar skills. We ordered these interchangeable chapters either by: (1) the output of our adaptive HMM personalization system (in the personalized condition), or (2) randomly from among the chapters the participant had not yet seen (in the non-personalized condition). We evaluate students before they participate in this story and afterwards with a fixed pre-test and post-test administered to both groups. These pre-tests and post-tests were disguised as chapters in the story and were administered by the robot, but were constant for both conditions. We evaluate the impact of our personalization system based on the differences in pre-test/post-test measures between groups.

6.3.1 Apparatus

In this experiment, each participant, ages 4 to 7, engaged in five one-on-one 20-minute long sessions with a small stationary robot named Keepon over the course of two weeks.

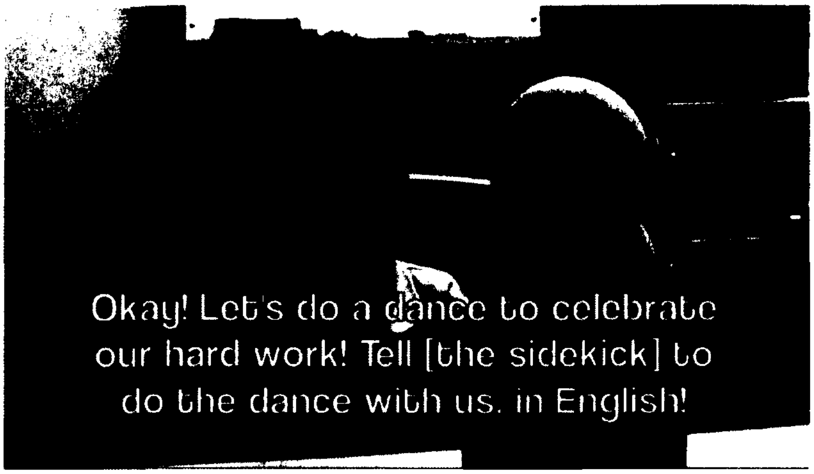

Okay! Let's do a dance to celebrate
our hard work! Tell [the sidekick] to
do the dance with us, in English!

FIGURE 6.1: A first grade student interacts with the robot tutor. The caption here is an English translation of what the robot is saying in Spanish. The robot told an adventure story to the participants, entirely in Spanish, and participants were asked to perform Spanish-to-English sentence translations to progress in the story. Participants performed between 30 and 40 translations per session, sessions lasting approximately twenty minutes. Each participant did five sessions over the course of two weeks.

6.3.2 Robot

The robot we used for this study, Keepon, is the same one we used in the studies in Chapters 2, 4, and 5. Keepon is an 11-inch tall, stationary, yellow, snowman-shaped robot with small, round eyes, one of which contains a camera, and a small, round nose containing a microphone. For a photograph, see **Figure 2.4**. In this study, the robot faced the participant and bounced while speaking in a personalized ordering of pre-recorded Spanish audio clips. See **Figure 6.1** below for the relative positioning of the robot and the participant.

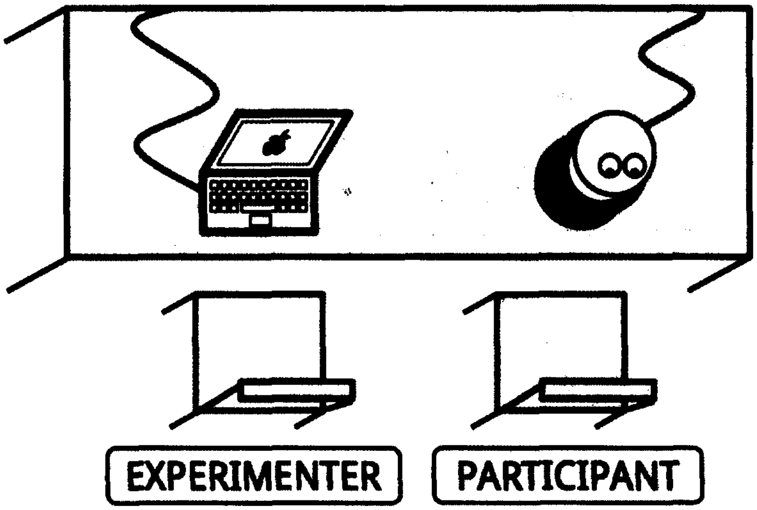


FIGURE 6.2: Overhead view of the experimental apparatus. The participant, a first grade student whose dominant language is Spanish, is seated facing the robot. The experimenter, who provides adult supervision and natural language processing for the robot, is seated beside the participant. The experimenter provided occasional encouragement and vocabulary assistance, as well as categorizing each of the participant's responses as either: correct, incorrect, irrelevant, or silent.

6.3.3 Participants

There were 19 participants in our study: 10 who received personalized lessons, and 9 who received non-personalized lessons. All of the participants were schoolchildren ages 4 to 7, attending the first grade. The participants were exclusively Spanish-dominant speakers, being raised in Spanish-dominant homes.

6.3.4 Experimenter

The participant was in the constant supervision of an adult during the course of this study. This adult, the present author, also played a role in the experiment. The experimenter and the participant sat side-by-side as seen in **Figure 6.2**. The experimenter performed three roles:

1. First and foremost, the experimenter monitored the safety and wellness of the child. There were no notable adverse incidents during the course of this study.

2. The second role of the experimenter was to provide natural language processing. We decided not to use Automated Speech Recognition (ASR) systems to process the participants' speech because such systems have relatively high error rates with children and non-native speakers (**Chen and Zechner 2011**; **Williams, Nix and Fairweather 2013**). Instead, the experimenter provided speech recognition information to the system by coding each of the participants' responses as either: 'correct', 'incorrect', 'irrelevant', or 'silent' using the objective rules described in **Section 6.4** below.

3. The last role of the experimenter was to provide occasional vocabulary assistance to participants in the study. The experimenter could only provide help with nouns, and not verbs, in order to preserve the integrity of the "make" vs. "do" distinction made entirely by participants.

“Make”	“Do”
M1 To construct or build. • make a cake • make dinner • make a bridge • make a tent • make a sound • make a decision	D1 To perform a job or activity. • do the dishes • do your homework • do a dance • do chores • do an assignment • do a project
M2 To elicit a reaction. • make him happy • make her smile/laugh • make it feel better • make us proud • make him pack • make sure that	D2 To perform unspecified action. • do something • do anything • do nothing • “What should we do?” • “Let’s do it!” • “How are you doing?”

TABLE 6.1: The English words “make” and “do” translate to one word in Spanish (“hacer”), and as a result many native speakers struggle to learn the distinction we make between these words when they learn English. We picked four such distinctions between these two English words, of which many more exist in the language. Every translation task participants did was designed to fit in exactly one of these categories.

6.4 Curriculum

During the course of this experiment, the robot engaged participants in an interactive adventure story task, a sample of which can be found in **Table 6.2**. In order to make progress through the story, participants were asked to translate between 30 and 40 sentences from Spanish to English per session. We used these translation tasks to teach four English grammar skills that are difficult for non-native speakers.

All of the translation tasks participants did in this study were sentences that, in English, contain either the words "make" or the word "do." In Spanish, both "make" and "do" translate to a single word, "hacer." As a result, native Spanish speakers often struggle to learn the distinction English speakers make between these words. Native Spanish-speakers often confuse the two. For example, children might say, "I made my homework," instead of "I did my homework," or "I did a goal in soccer today," instead of, "I made a goal."

In the English language, there are as many as ten distinct categories of usage for these two words that distinguish them from one another, depending on the ESL curriculum one chooses. For this work, we chose just four of these categories, two for the word "make" and two for "do." All of our translation tasks fit exactly into one of these four categories, as described in **Table 6.1**. We chose to teach four categories rather than teaching all ten so as to ensure that there were enough observations per participant per category to train our model in the allotted time for the study. We treat each of these four category as a distinct skill in the model.

Each translation that the participants did was interpreted by the experimenter, whose role is described in **Section 6.3.4** above. The experimenter categorized each of the participants' translation tasks using the following set of objective rules. Correctness in the context of this study was determined entirely by the verb used in the translation. When participants used the correct verb (either "make" or "do") the translation was marked 'correct,' regardless of the rest of the translation. If the participant used the verb "do" in the place of "make" or vice versa, the translation was marked 'incorrect.' If neither verb was used in the translation, it was marked 'irrelevant.' If the participant did not respond, 'silent' was marked.

6.4.1 Sessions

There were five total sessions with each participant, no more than one per day, held over the course of two weeks, each lasting approximately twenty minutes. The sessions were conducted as follows:

- The first session was a pre-test. Its contents were identical for participants in both groups. There were 40 translation tasks in this session, 10 per skill.
- The second, third, and forth sessions consisted of 30 translation tasks each. These middle sessions were composed of 3 interchangeable chapters each, with 10 translation tasks per chapter. Each chapter targets exactly one of the four grammatical skills described above. The bundling of 10 translation tasks into each interchangeable chapter limited the flexibility of our personalization system but was a necessary

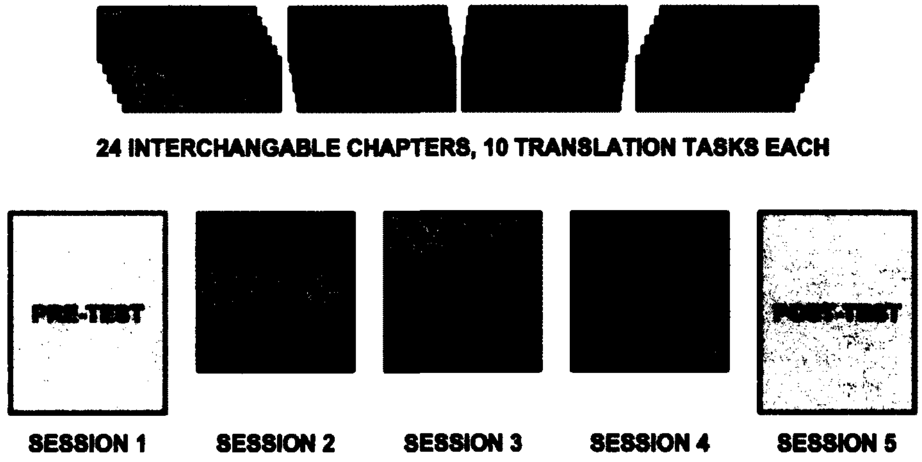


FIGURE 6.3: Participants engaged in five sessions over the course of a two week period. The first session was a pre-test, the same across all participants, with ten translation tasks per skill. The middle three sessions were comprised of 3 interchangeable chapters, each focussed on one specific skill, and each containing 10 translation tasks. The post-test was the same across all conditions and contained 10 translation tasks per skill. The ordering of the interchangeable chapters varied based on the condition, as described is Section 6.4.2 below.

tradeoff to keep our target population (4-7 year olds) engaged in a multi-day learning task. We authored a total of 24 interchangeable chapters for this study, 6 that targeted each of the 4 skills. In total, each participant saw only 9 of the 24 interchangeable chapters. Again this limitation was necessitated by the population, in order to avoid fatigue. For a visual representation of the content of each session see Figure 6.3. The ordering and selection of the lessons was determined by the condition the participant was in:

- In the personalized lessons condition, the episodes were ordered based on a Hidden Markov Model (HMM) that we built for each participant and skill.

The model for each skill consisted of three hidden states, either (1) the participant does not know the skill, (2) the participant does know the skill, or (3) the participant has forgotten the skill. In the personalized condition, the lessons targeting 'not-known' skills were chosen first, among those that the participant had not already seen. The parameters of the HMM were updated after the pre-test and then again after each interchangeable chapter. The details of the model can be found in **Section 6.4.2** below. If no skills were 'not-known', then lessons that targeted 'forgotten' skills were chosen randomly among those not yet seen. Lastly, if all of the skills were 'known', the tutor chose a random episode among the ones the participant had not yet seen.

  - In the non-personalized lessons condition, participants received a random episode that they had not yet seen, distributed uniformly over the 4 skills. Because participants saw 9 total chapters, they saw one skill three times and the others twice. This condition is meant to simulate group classroom instruction in that the lessons are not in an order best suited to any particular student, but rather evenly sampled across all the material at the teacher's discretion.

- The fifth and last session of the study was a post-test. Like the pre-test, there were 40 translations, ten per skill, and every participant saw the same content in their fifth session with the robot, regardless of their group. We compare the results of the pre-test and post-test scores across groups in **Section 6.6** below.

Robot Tutor Prompt (Translated to English)	**Skill**
It's so beautiful here! There are so many mountains. I think that's a cave over there! Does that look like a cave?	—
Let's get closer and see! I've never seen a cave before. «HAPPY BARK» I think it IS a cave! That's so cool! Let's go explore. We shouldn't spend too much time here, though, since we have a lot to do today! Please tell Toby, in English, **to make sure we leave soon.**	M2
«HAPPY BARK» Alright, let's go inside! It's kind of dark in here. I hope we don't get lost! Maybe we should make a map. Will you ask Toby, in English, if he knows how to **make a map?**	M1
«CONFUSED BARK» I'm not sure that he knows how to make a map. Please tell him in English to first **make a picture of the cave.**	M1
«HAPPY BARK» Okay, thank you! We should make sure the picture is big enough for us to see, though. Please tell him in English to **make a BIG map.**	M1
«HAPPY BARK» Great, he's making the map! Now we won't get lost. We should also make notes of what we see, so that I can tell my friends when I get back home! Please tell Toby, in English, that we should **make notes.**	M1

TABLE 6.2: Sample of the robot's dialogue targeted at skill 'M1.' The robot's dialogue was pre-scripted and pre-recorded in Spanish, the English translation of which is presented above. The bolded portions are the translation tasks that participants were asked to perform. 'Toby', above, refers to an imaginary dog character that only understands English commands. For more information about the dialogue, see **Section 6.4**. See **Table 6.3** below for the original Spanish dialogue.

Robot Tutor Prompt (Original Spanish)	Skill
Es muy hermoso aquí. Hay muchas montañas. Creo que es la cueva allá. ¿Esos te parece una cueva?	—
Hay que cercarnos más y ver. Yo nunca he visto una cueva antes. «LADRA CONTENTA» Creo que sí es una cueva. ¡Eso es muy padre! Hay que explorar. Pero no deberíamos gastar tanto tiempo aquí, como tenemos mucho que hacer hoy. Por favor di la Toby, en inglés, **que se segura de que nos vayamos pronto.**	M2
«LADRA CONTENTA» Bueno, hay que ir adentro. Está un poco oscuro aquí. Espero que no los perdamos. Tal vez deberíamos hacer un mapa. ¿Le preguntes a Toby, en inglés, si él sabe cómo **hacer un mapa**?	M1
«LADRA CONFUNDIDA» Yo no estoy segura que él sabe cómo hacer un mapa. Por favor, dile en inglés que primero a **don un dibujo de la cueva.**	M1
«LADRA CONTENTA» Bueno, gracias. Pero deberíamos hacer el dibujo lo suficientemente grande para poderlo ver. Por favor, dila en inglés que **hago una mapa grande.**	M1
«LADRA CONTENTA» Genial. Él está haciendo el mapa. Ahora no los perderemos. Tan bien deberíamos hacer notas de lo que miramos. Así le puedo decir a mis amigos cuando regrese a casa. Por favor dile a Toby, en inglés, que deberíamos **hacer notas.**	M1

TABLE 6.3: Sample of the robot's dialogue targeted at skill 'M1', in the original Spanish language, as spoken by the robot. See Table 6.2 above for English translation.

6.4.2 Personalization

There were two conditions in this study, personalized lessons and non-personalized lessons. We discuss the personalization condition below; for more information about the non-personalized condition please see Section 6.4.1 above.

The goal of the personalization in this system is to sequence the interchangeable chapters we wrote to best suit the skill competencies of an individual student, by challenging him or her with the translation tasks that he or she needs to practice most. Here we describe a system that takes as input the series of translation task observations coded by the experimenter, as described in **Section 6.3.4**, and produces as output one of the four skills, by which the robot chose the next interchangeable chapter to give participants in the personalized lessons condition.

For each skill and each participant, we created independent same-structured Hidden Markov Models (HMMs) with three hidden states: (1) the participant does not know that skill, (2) the participant does know that skill, or (3) the participant forgot that skill. To see how these states are connected, see **Figure 6.4**.

There were four observable states in this model: (1) a correct answer, (2) incorrect answer, (3) irrelevant answer, or (4) no answer. For more information about how these observable states were recorded by the experimenter, see **Section 6.3.4**.

For each skill, the model was trained on the subset of the translation tasks targeting that skill alone. Because each translation task targeted exactly one of the four available skills, each of the four HMMs was trained on approximately one fourth of the collected data across all participants.

We fixed some parameters of the HMM in advance, and learned the rest with the Baum-Welch algorithm based on the collected data (**Welch 2003**). In total, we fixed 4 parameters, and learned the remaining 14. The learned parameters were first learned based on the pre-test data and then updated with each new chapter's worth of data as it was collected.

We fixed the initial distributions of the hidden states for all four skills, based on the expert estimate of an ESL educator. She estimated that:

$$P(\text{KNOWS-SKILL}) = 0.2,$$

$$P(\text{FORGOT-SKILL}) = 0.4, \text{ and}$$

$$P(\text{DOES-NOT-KNOW-SKILL}) = 0.4.$$

We also fixed the transition probability that a participant gives a correct answer given that he or she is in the 'KNOWS-SKILL' state. This choice was inspired by mastery learning literature in education research, in which students are expected to demonstrate mastery of a skill before learning another (Kulik, Kulik and Bangert-Drowns 1990). In this model, we wanted to ensure that the transition from 'KNOWS-SKILL' to a 'CORRECT' answer was not learned by the Baum-Welch algorithm as a relatively low probability, thereby overestimating the competency of participants. Instead, we set a relatively high requirement for the HMM to end up in the 'KNOWS-SKILL' hidden state by setting $P(\text{CORRECT}|\text{KNOWS}) = 0.9$ for all four skills.

We apply the Viterbi algorithm to pick the most likely hidden state given the series of observations (Forney Jr. 1973). This tells us which of the four skills each student knows, doesn't know, or has forgotten, and we use that information to choose a personalized lesson for each participant as follows:

- If any skill is unknown, the robot chose a random lesson targeting one of those skills from among the lessons that the participant had not yet seen.
- If any skill is forgotten, the robot chose a random lesson targeting one of those skills from among the lessons that the participant had not yet seen.
- If no skills are unknown and no skills are forgotten, then all skills are known and we choose a random lesson targeting any skill from among the lessons that the participant had not already seen.

The aim of this personalization is to target unknown or poorly understood skills first. Though this challenges students, it enables them to distinguish skills from one another more accurately. As students learn the patterns inherent to each skill, they start to improve across all skills.

Our model includes a hidden state for forgetting a skill as result of our experience running this experiment with a pilot group over the course of five weeks. We noted that participants' performance worsened between sessions, especially sessions that had more than a week-long gap between them. This internal state is likely not necessary for shorter-term automated personalization systems.

We compare how this personalization system affected student learning gains relative to a non-personalized control group in Section 6.6 below.

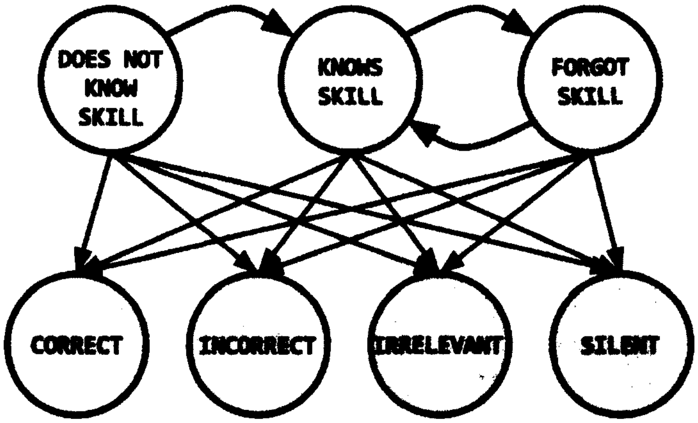


FIGURE 6.4: The Hidden Markov Model (HMM) used to sequence curriculum for the personalized group. Four simultaneous copies of this model were trained and run for each student, one for each of the English grammar skills defined above. Implementation details of the HMM can be found in Section 6.4.2.

6.5 Procedure

Participants were divided into two experimental conditions but the sole difference between groups was the ordering of the translation tasks in the second, third, and forth sessions. The participants were blind to the condition they experienced. All participants followed the same procedure in this study as outlined below.

Before the experiment began, a voluntary consent form was sent to parents of potential participants, all of whom were in the same first grade class in a bilingual school, with help from school administrators. Students whose parents consented were informed that they could stop their participation in the study at any time, for any reason, simply by walking away from the robot. Participants were supervised during the course of the

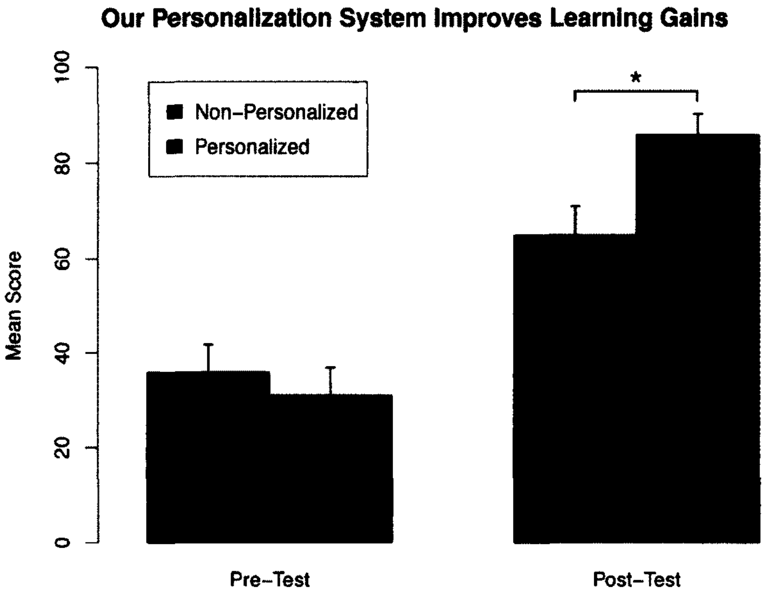


FIGURE 6.5: Pre-test and post-test results across experimental groups, indicating the effectiveness of our personalization system. Participants who received personalized lessons performed significantly better on the post-test ($M = 84, SD = 8$) than participants who received non-personalized lessons ($M = 63, SD = 9$), $p(16) < 0.03$. Error bars depict standard error. Statistical significant determined by unpaired Student's T-Tests.

study by the experimenter.

Participants engaged in five sessions of approximately twenty minutes in length, no more than once per day, over the course of two weeks. We present the results of these sessions below.

6.6 Results

We investigated the effects of our personalization system on a longer-term robot tutoring interaction. Participants performed 30-40 translation tasks per session and an experimenter coded each translation as either: 'correct', 'incorrect', 'irrelevant', or 'silent'. We compare the mean percentage of 'correct' answers between groups to evaluate our personalization system.

Participants who received personalized lessons (n=10) performed significantly better on the post-test ($M = 84, SD = 8$) than participants who received non-personalized lessons (n=9) ($M = 63, SD = 9$), $p(16) < 0.03$. Statistical significant determined by unpaired Student's T-Tests. This result indicates that our personalization system led to significantly increased learning gains, by a mean of 2.0 standard deviations, corresponding to an improvement in the 98th percentile of scores in the non-personalized group.

There was no significant difference in the pre-test scores between these two groups, with mean scores of ($M = 38, SD = 11$) for the non-personalized group and ($M = 36, SD = 13$) for the personalized group. This result indicates that the two groups started with roughly the same knowledge and, as a result of the personalization system, the group that received personalized lessons learned significantly more over the course of the study than the group that received non-personalized lessons.

Another result of our personalization system is the differences in correctness scores between groups during the second session, which was either the first personalized lessons session for the personalized group, or the first non-personalized lessons session for the

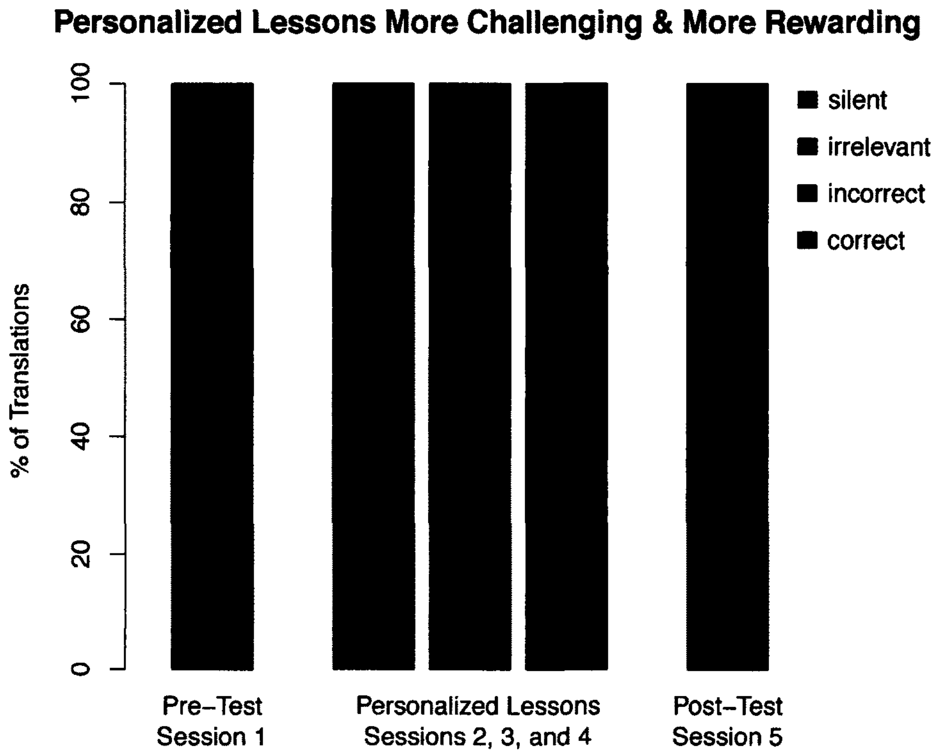


FIGURE 6.6: Distribution of answers given by participants in the personalized lessons condition across all five sessions with the robot. Between the pre-test and post-test, we see a momentary drop in correctness scores that then seems to increase exponentially until the post-test. This result validates our main manipulation, in which we were attempting to challenge students to the hardest problems first. The lower initial scores, rising sharply over time, indicate that our personalization system correctly identified which skills each participant needed more practice with and that the personalized lessons each participant received caused a sharp increase in learning gains over time.

non-personalized group. These data are plotted in **Figure 6.6** and **Figure 6.7**. The mean percentage of correct answers was significantly lower in the personalized lessons group,

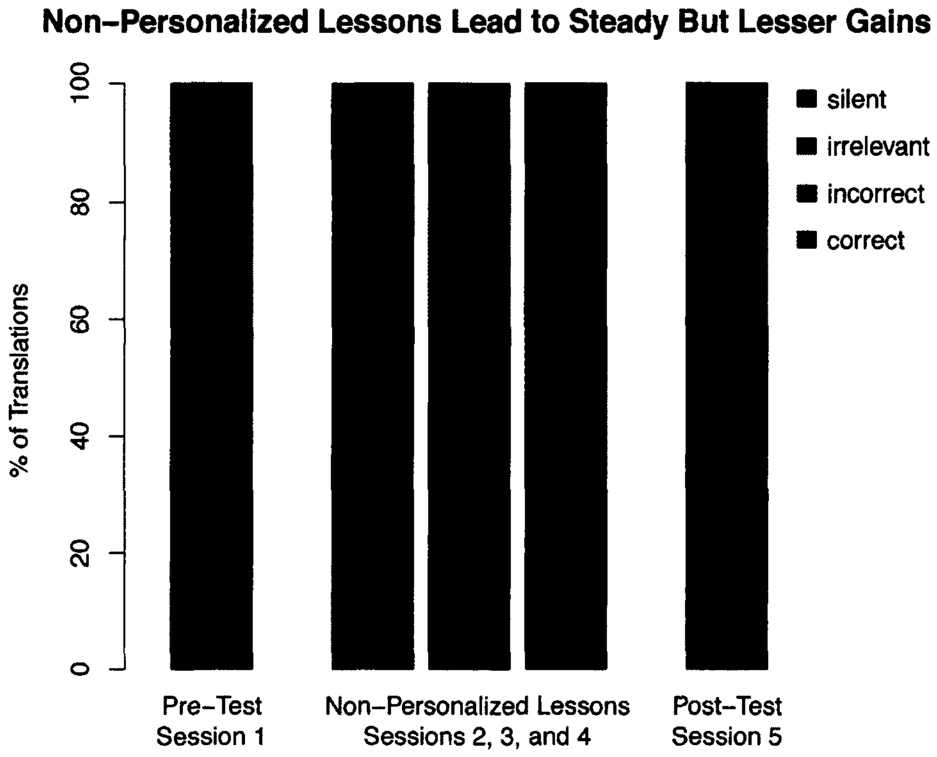


FIGURE 6.7: Distribution of answers given by participants in the non-personalized lessons condition across all five sessions with the robot. Between the pre-test and post-test, we see a steady increase in correct responses, and corresponding decrease in the incorrect, irrelevant, and silent responses. This result is consistent with the expectation of a typical classroom learning experience, in which we expect students to perform incrementally better the more material they are exposed to. Some students may be bored while others may be failing, but the mean continues to rise.

($M = 28, SD = 8$), than in the non-personalized group, $M(50, SD = 10)$, $p(14) < 0.01$. This result indicates that participants who received personalized lessons found the lessons more challenging than those who received non-personalized lessons. We can conclude from this that our personalization system correctly identifies the skills in which each participant lacks competency, and can be used successfully to sequence curriculum in order to challenge students.

6.7 Discussion

The most significant result in this study is the extent to which personalization impacted learning gains. However, even participants who received non-personalized lessons significantly improved their knowledge during the course of this study. Participants receiving non-personalized lessons improved their scores by an average of 25 points ($M = 25\%, SD = 14\%$) between pre-test and post-test. This is evidence that simply the act of repeated practice, with a robot, is enough to stimulate significant learning gains in an ESL domain. When personalization is added to a robot language tutor, which would be useful on its own, the gains are even higher.

In the data we collected, we can see a difference in the patterns of the correctness data between conditions over the course of the five sessions. In **Figure 6.6**, where the personalized participants response distributions are plotted, there is an exponential-like growth in the 'correct' answers. The personalized lessons caused participants to struggle with harder problems in the second session with the robot and thus made the rest of their time significantly more effective. Though they faced material that was more personally challenging, and thus failed more often early in the study, their post-test scores were very high. Whereas, in **Figure 6.7**, which shows the corresponding data for the participants who received non-personalized lessons, we see a growth pattern that is close to linear rather than exponential. This may reflect a classroom style educational experience, in which curriculum is sequenced by a teacher to suit the majority of the class rather than any individual, and as a result, produces steady learning gains that are not as quick as with a personal tutor. The patterns in these data clearly favor the personalized model, but only if the initial challenge presented by personalization is not overwhelming to the point of frustration on the part of students. So long as students stay with the tutoring, they will achieve much better end results.

Another interesting outcome of this study is the relative scarcity of 'incorrect' answers among the data in either group. The mean overall occurrence of 'incorrect' answers across both groups and all sessions was only 7% ($SD = 3\%$). For a visual representation of the distribution of these answers per session and group, see **Figure 6.6** and **Figure 6.7**. As a reminder, 'incorrect' answers are those where a participant used "make" in a sentence that was intended to be translated as "do" or vice versa. The rarity of this phenomena in our study is not necessarily representative of its frequency in the participants' natural conversations, where less pressure and structure is applied to their English language speech. Mistakes are likely more common when there is no study being run. Whereas, in the context of this study, it is likely that the relative rarity of incorrectness reflects a self-awareness on the part of the participants of their own lack of understanding of this grammatical distinction in English. Participants commonly avoided making the distinction between "make" and "do" by omitting the verb in their sentences or changing the meaning of the sentence slightly to avoid having to make the distinction altogether.

These instances are what make up the 'irrelevant' data in the above plots. The relative abundance of 'irrelevant' and even 'silent' responses compared to 'incorrect' also seems to corroborate the claim that participants avoided making the distinction when they knew they did not know the skills well enough. Toward the end of the study, this behavior was reduced, likely because participants had more knowledge of the skills.

The relatively large increase in skill competency, across both groups, as measured by the post-test, raises the question of whether these skills can be transferred to students' daily speech and ESL class performance. Though this is not the research question we ask in this work, as we are focused on creating effective personalization systems for robot tutors, it is a question one should ask of any education intervention in the long term. Do robot tutoring interventions like the one we made produce learning gains that transfer into daily life? All of the participants in this study were students in the same first grade class and their teacher commented on an improvement in the days after our work without the authors' prompting. It was our experience that, generally, students were enthusiastic about interacting with the robot, even despite its limited capabilities at present, and that, likely, the skills improved in the study did transfer, at least to some small extent, to the students' lives. As such systems become more robust, researchers may want to perform followup work to see what the long-term impacts are of such interventions.

6.8 Conclusion

In this work, we describe a personalization system for longer-term robot tutoring and we test our system with English as a Second Langauge (ESL) curriculum targeted towards Spanish-dominant first grade students. In this study, participants were divided into one of two conditions: they either received personalized lessons as decided by our personalization system, or they received non-personalized lessons chosen at random but evenly distributed among the ESL skills we targeted. We found that the participants who received personalized lessons significantly outperformed participants who received non-personalized lessons by a factor of 33%. We also found evidence that our personalization system correctly identifies a students weakest skills and can be used to sequence curriculum to maximize a robot tutor's effectiveness.

Chapter 7

Conclusion

This dissertation makes three contributions to the study of personalization in robot tutoring: (1) we provide evidence for improved student learning gains associated with the physical presence of a robot tutor, (2) we deliver experimentally-derived design guidelines for future work in robot tutoring, and (3) we provide novel robot tutoring personalization systems and demonstrate that these systems improve student learning outcomes over non-personalized systems by 1.2 to 2.0 standard deviations, corresponding to gains in the 88th to 98th percentile.

We summarize the key details of our contributions below and discuss direction for our future work.

7.1 Embodiment affects learning gains

In **Chapter 2** we investigated to what extent the physical presence of a robot tutor can affect student learning outcomes. We conducted an experiment with three conditions, in which the instructional content was the same across all three conditions but the content was delivered by either: (1) a physically-embodied robot tutor, (2) an on-screen character tutor, based on video footage of the robot in the first condition, or (3) a voice-only tutor with no physical or virtual embodiment, which used the same voice as in the previous two conditions. We measured how long participants took to complete four logic puzzles called 'Nonograms.' The fourth Nonogram puzzle and the first Nonogram puzzle consisted of the same gameboard, disguised by a 90° rotation, thereby allowing us to track each individual participant's puzzle-solving skills from the first time they need to applied the skills necessary to solve that puzzle to the last time they needed to apply exactly those same skills to solve the final puzzle.

We found that participants who received lessons from a physically-embodied robot tutor significantly outperformed participants who received the same lessons from an on-screen video representation of that robot, as well as significantly outperforming the participants who received the same lessons from a voice-only tutor. Participants who received tutoring from the physical robot improved their same-puzzle solving time significantly more than participants in the other two groups. Our data shows that the physical presence of a robot tutor can lead to learning gains of 0.3 standard deviations, corresponding to gains in the 62nd percentile. The survey data, however, did not conclusively reveal the underlying mechanism for these learning gains. The data did indicate that participants were less "annoyed/distracted" by the physically-embodied tutor than the other tutors. Perhaps then the underlying reason for their learning gains was an increase in attention and engagement with the tutor. Our work is the first to investigate the effect of the physical presence of a robot tutor and, as such, more work is needed to assess the underlying mechanisms of the effect we found.

7.2 Experimentally-driven design guidelines point to the importance of affect in robot tutoring interactions

In **Chapter 3** and **4** we used robots as students to investigate how human tutors perform personalization. We use the results from these studies to provide experimentally-derived design guidelines for future work in automated personalization systems.

In the first of these studies, described in **Chapter 3**, we investigated how human tutors personalize their instruction when they teach robot students of differing abilities. Each participant in this study tutored two robot students, first teaching one robot student then teaching the other robot student. One of the robot students was significantly more successful in the learning tasks than the other. Participants were led to believe that the robots were learning based on the verbal instruction participants were encouraged to give, however, the actions of each of the robots was planned ahead of time and constant across all participants. This manipulation allowed us to compare how human tutors taught these two kinds of robot students differently. We measured the quantity, timing, and affective quality of the participants vocalizations and compared how participants personalized their instruction between the more successful robot student and the less successful robot student.

Our results indicate that human tutors treat students of differing abilities vastly differently. The most significant differences were that: (1) participants talked progressively less to the more successful student during the course of this study, likely due to an inference that the more successful student needed less scaffolded support, (2) participants used more strongly affective speech towards the less successful student, most of which was rated as positive, such as encouragement and motivational vocalizations, and (3) participants provided more guidance to the less successful robot student in the final trial than they did on the same trial for the more successful student, indicating that they expected the less successful student to fail more often than the more successful student. Our results indicate that the quantity and affective quality of a tutor's speech is heavily personalized based on an individual student's ability level. Few automated tutoring systems vary the emotional content or the amount of tutoring offered to students to the same degree as we found human tutors do.

In our second study employing robots as students, described in **Chapter 4**, we investigated to what extent human tutors personalize their teaching to robot students that behave identically in all learning tasks but have differing personalities. Participants in this study taught a robot student, and each time the student completed a task, it would respond with either: (1) emotionally appropriate responses, (2) often emotionally-inappropriate responses, or (3) apathetic responses. In this study, as in the previous, we led participants to believe that the robot was learning from the verbal instructions and, in this case, physical demonstrations that we encouraged participants to give the robot. In this study, participants were allowed to provide as many demonstrations as they wished to the robot students. We measured how many demonstrations each group elected to do and we also measured the accuracy of each demonstration.

We found that, even when students perform identically across all learning tasks, human tutors still personalize their instruction based on a student's emotional responses. Our data show that participants who taught the robot student who gave emotionally-appropriate responses performed significantly more demonstrations and performed significantly more accurate demonstrations on average than participants who taught the apathetic or participants who taught the often emotionally-inappropriate robot students. Between participants who taught the apathetic robot student and participants who taught the often emotionally-inappropriate robot student there were no significant differences in their behavior or survey ratings of the robot, indicating that students who seem apathetic produce the same disengagement in human tutors as students who actively misbehave. We also found that human tutors perceived the robot student who gave emotionally-appropriate responses as significantly smarter than participants perceived the other two students to be, despite the fact that the behavior of the robot students was identical across all three conditions. From these results we propose two guidelines for future work in personalization of automated tutors: to be more human-like, (1) automated tutors should respond differently to students who produce differing patterns of emotional responses and (2) automated tutors should detect when student engagement is so low as to affect the perception of their performance on learning tasks and intervene by offering a break or a reward instead of continuing to tutor to students who are disengaged. Some automated tutoring systems do model students' affective states but few systems produce the same degree of personalization that we found human tutors do.

7.3 Personalization systems for robot tutoring dramatically increase learning gains

In our final two chapters, we contribute novel automated personalization systems for robot tutors: two systems for shorter-term robot tutoring interactions and one system for longer-term interactions.

As the first research group to investigate the role of personalization in robot tutoring, we were interested in establishing a minimum threshold for the effects of personalization on student outcomes. In Chapter 5 we investigate to what extent the personalization of a robot tutor can affect student learning gains over the course just a single tutoring session. In this work we designed an experiment with four conditions: (1) a condition in which participants received personalized lessons from a robot tutor based on our first personalization system, a simple additive model of the participant's skills, (2) a condition in which participants received personalized lessons from a robot tutor based on our second personalization system, a slightly more sophisticated Bayesian model, (3) a condition in which participants received non-personalized lessons from the same robot tutor as in the first two conditions, and (4) a condition in which participants were asked to perform the same learning tasks as in the previous three conditions but with no lessons or tutoring whatsoever.

We find that personalization has a significant impact on student learning outcomes in robot tutoring, even in the course of a single session with the robot tutor. Our results show that the additive model produced learning gains of 1.0 standard deviations, corresponding to gains in the 85th percentile, and that the Bayesian model produced significantly higher learning gains of 1.4 standard deviations, corresponding to gains in the 92nd percentile. Participants in both of these groups performed significantly better than students in either of the two control groups, those that received non-personalized lessons and the ones that received no lessons. We also found that participants rated both personalized tutors as significantly more "smart" and significantly less "annoying" than the non-personalized condition, reaffirming our findings that even in the course of a single session with a robot tutor, personalization can significantly impact the perception of the interaction by students and the learning gains those students ultimately make.

In our final chapter, **Chapter 6**, we created an automated personalization system intended for longer-term interactions and tested its effectiveness in a two-week-long five-session interaction with children. Our system uses a Hidden Markov Model (HMM) with three hidden states and four observation states, in which fourteen of the transition probabilities were learned on-line during the course of the interaction. To investigate the effectiveness of this system we designed an experiment with two conditions: either (1) participants received personalized lessons from a robot tutor based on the HMM or (2) participants received non-personalized lessons from the same robot tutor. We measured the number of questions participants answered correctly in a fixed pre-test and post-test, where keeping the pre-test and post-test constant between groups allowing us to compare performance between as well as within group.

We found that participants who received personalized tutoring based on our model significantly outperformed students who received non-personalized tutoring by an average of 2.0 standard deviations, corresponding to gains in the 98th percentile. This result confirms the effectiveness of our longer-term personalization system. We also found that participants in both groups learned a significant amount of the content over the course of the study: students who received non-personalized tutoring achieving a mean post-test score roughly double their pre-test score, whereas participants who received personalized tutoring nearly tripled their pre-test scores. The participants who received non-personalized lessons made steady but slow gains, whereas the participants who received personalized lessons were more challenged and made more drastic gains. We conclude that our personalization system correctly identified the skills with which individual students needed the most practice and challenged students with personalized lessons that resulted in significantly better student outcomes over a similar non-personalized system.

7.4 Future Work

In summation, this dissertation attempts to lay some of the groundwork necessary to produce personalized robot tutors that serve as an at-home interaction partner to keep students engaged and on track in their homework. We know from our work that the physical presence of robots give them a unique ability to keep students engaged, and so we have shown that such homework helper robots have the potential to be more successful than on-screen tutors. We know from our work that, in developing these tutors, its important to note the affective communication, both on the tutor's side and the student's. Last, we know that the algorithms we developed produce significantly improved learning gains and can act as a starting point for future work. The next steps towards our goal are to attempt longer-term studies and the develop appropriate machine learning algorithms that can model a student's knowledge state during weeks-long, months-long, or years-long tutoring.

Bibliography

Argall, B. D., S. Chernova, M. Veloso and B. Browning. 2009. "A survey of robot learning from demonstration." *Robotics and Autonomous Systems* 57(5):469 – 483.

Aronson, E. 1969. "The Theory of Cognitive Dissonance: A Current Perspective." *Advances in Experimental Social Psychology* 4:1–34.

Auerbach, E. R. 1993. "Reexamining English only in the ESL classroom." *Tesol Quarterly* 27(1):9–32.

Bainbridge, W.A., J. Hart, E.S. Kim and B. Scassellati. 2008. "The effect of presence on human-robot interaction." *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on* pp. 701–706.

Baker, R. S., A. T. Corbett and K. R. Koedinger. 2004. Detecting student misuse of intelligent tutoring systems. In *Intelligent tutoring systems.* Springer pp. 531–540.

Baker, R. S., A. T. Corbett and V. Aleven. 2008. More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In *9th International Conference on Intelligent Tutoring Systems.* ITS '08 Berlin, Heidelberg: Springer-Verlag pp. 406–415.

Bartelmus, C. 2008. LIRC - Linux Infrared Remote Control. In *http://www.lirc.org/*.

Baylor, A. and S. Ebbers. 2003. The Pedagogical Agent Split-Persona Effect: When Two Agents are Better than One. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2003*, ed. David Lassner and Carmel McNaught. Honolulu, Hawaii, USA: AACE pp. 459–462.

Bloom, B. S. 1984. "The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring." *Educational Researcher* 13(6):4–16.

Cakmak, M. and A. L. Thomaz. 2012. Designing robot learners that ask good questions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM pp. 17–24.

Callahan, R. M. 2005. "Tracking and high school English learners: Limiting opportunity to learn." *American Educational Research Journal* 42(2):305–328.

Chen, M. and K. Zechner. 2011. Computing and Evaluating Syntactic Complexity Features for Automated Scoring of Spontaneous Non-native Speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11 Stroudsburg, PA, USA: Association for Computational Linguistics pp. 722–731.

URL: *http://dl.acm.org/citation.cfm?id=2002472.2002564*

Chi, M. T. H., N. Leeuw, M. Chiu and C. LaVancher. 1994. "Eliciting self-explanations improves understanding." *Cognitive science* 18(3):439–477.

Chi, M. T. H., S. A. Siler, H. Jeong, T. Yamauchi and R. G. Hausmann. 2001. "Learning from human tutoring." *Cognitive Science* 25(4):471–533.

Cohen, J. 1968. "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." *Psychological Bulletin* 70(4):213–220.

Cohen, P. A., J. A. Kulik and C. L. C. Kulik. 1982. "Educational outcomes of tutoring: A meta-analysis of findings." *American educational research journal* 19(2):237–248.

Conati, C. and H. Maclaren. 2009. "Empirically building and evaluating a probabilistic model of user affect." *User Modeling and User-Adapted Interaction* 19(3):267–303.

Craig, S., A. Graesser, J. Sullins and B. Gholson. 2004. "Affect and learning: An exploratory look into the role of affect in learning with AutoTutor." *Journal of Educational Media* 29(3):241–250.

Creed, C. and R. Beale. 2008. "Psychological Responses to Simulated Displays of Mismatched Emotional Expressions." *Interacting with Computers* 20(2):225–239.

Dahlbäck, N., A. Jönsson and L. Ahrenberg. 1993. Wizard of Oz Studies: Why and How. In *Proceedings of the 1st International Conference on Intelligent User Unterfaces*. ACM pp. 193–200.

Desmarais, M. C. and R. S. Baker. 2012. "A Review of Recent Advances in Learner and Skill Modeling in Intelligent Learning Environments." *User Modeling and User-Adapted Interaction* 22(1-2):9–38.

Devedzic, Vladan and John Debenham. 1998. An Intelligent Tutoring System for Teaching Formal Languages. In *Intelligent Tutoring Systems*, ed. B. P. Goettl, H. M. Halff,

C. L. Redfield and V. J. Shute. Vol. 1452 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 514–523.

D'Mello, S. 2012. Monitoring affective trajectories during complex learning. In *Encyclopedia of the Sciences of Learning.* Springer pp. 2325–2328.

D'Mello, S. K., S. D. Craig, B. Gholson, S. Franklin, R. W. Picard and A. C. Graesser. 2005. Integrating affect sensors in an intelligent tutoring system. In *Affective Interactions: The Computer in the Affective Loop Workshop at.* pp. 7–13.

DogsBody & Ratchet Software. 2009. MySkit – Performance Editor for PLEO. In *http://www.dogsbodynet.com/myskit/index.html.*

Eskenazi, M. 2009. "An overview of spoken language technology for education." *Speech Communication* 51(10):832–844.

Forney Jr., G. D. 1973. "The Viterbi algorithm." *Proceedings of the IEEE* 61(3):268–278.

Gold, B., N. Morgan and D. Ellis. 2011. *Speech-Recognition Overview.* John Wiley & Sons, Inc. pp. 59–69.

Graesser, Arthur C., Vasile Rus, Sidney K. D'Mello, G. T. Jackson, D.H. Robinson and G Schraw. 2008. "AutoTutor: Learning through natural language dialogue that adapts to the cognitive and affective states of the learner." *Recent innovations in educational technology that facilitate student learning* pp. 95–125.

Harmonix. 2005. "Guitar Hero." PlayStation.

Hogan, K. E. and M. E. Pressley. 1997. *Scaffolding student learning: Instructional approaches and issues.* Brookline Books.

Humes, K., N. A. Jones and R. R. Ramirez. 2011. *Overview of race and Hispanic origin, 2010.* U.S. Department of Commerce.

Hyun, E., H. Yoon and S. Son. 2010. "Relationships Between User Experiences and Children's Perceptions of the Education Robot." *5th ACM/IEEE International Conference on Human-Robot Interaction* pp. 199–200.

IguanaWorks. 2008. IguanaWorks. In *http://iguanaworks.net/.*

Johnson, J. S. and E. L. Newport. 1989. "Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language." *Cognitive Psychology* 21(1):60–99.

Kidd, C. D. and C. Breazeal. 2004. "Effect of a robot on user perceptions." *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on* 4:3559–3564.

Kidd, C. D. and C. Breazeal. 2008. "Robots at home: Understanding long-term human-robot interaction." *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems* pp. 3230–3235.

Kiesler, S., A. Powers, S.R. Fussell and C. Torrey. 2008. "Anthropomorphic interactions with a robot and robot-like agent." *Social Cognition* 26(2):169–181.

Kim, E. S., D. Leyzberg, K. M. Tsui and B. Scassellati. 2009. How people talk when teaching a robot. In *HRI '09: Proceedings of the 4th ACM/IEEE international conference on Human robot interaction.* New York, NY, USA: ACM pp. 23–30.

Konami. 1998. "Dance Dance Revolution." Arcade System.

Kozima, H., C. Nakagawa and Y. Yasuda. 2005. Interactive robots for communication-care: a case-study in autism therapy. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on.* pp. 341–346.

Kulik, C. C., J. A. Kulik and R. L. Bangert-Drowns. 1990. "Effectiveness of mastery learning programs: A meta-analysis." *Review of educational research* 60(2):265–299.

Lee, J. K., R. L. Toscano, W. D. Stiehl and C. Breazeal. 2008. "The design of a semi-autonomous robot avatar for family communication and education." *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on* pp. 166 –173.

Lee, M. K., J. Forlizzi, S. B. Kiesler, P. E. Rybski, J. Antanitis and S. Savetsila. 2012. "Personalization in HRI: a longitudinal field experiment." *7th ACM/IEEE International Conference on Human-Robot Interaction* pp. 319–326.

Lehman, Blair, Melanie Matthews, Sidney D'Mello and Natalie Person. 2008. What Are You Feeling? Investigating Student Affective States During Expert Human Tutoring Sessions. In *Intelligent Tutoring Systems*, ed. B. P. Woolf, E. Aïmeur, R. Nkambou and S. Lajoie. Vol. 5091 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 50–59.

Leite, Iolanda, Ginevra Castellano, André Pereira, Carlos Martinho and Ana Paiva. 2012. "Long-Term Interactions with Empathic Robots: Evaluating Perceived Support in Children." *Lecture Notes in Computer Science* 7621:298–307.

Lester, J. C., S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone and R. S. Bhogal. 1997. "The persona effect: affective impact of animated pedagogical agents." *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* pp. 359–366.

Levy, M. 1997. *Computer-Assisted Language Learning: Context and Conceptualization.* ERIC.

Levy, M. and G. Stockwell. 2013. *CALL dimensions: Options and issues in computer-assisted language learning.* Routledge.

Littleton, K. and P. Light. 1999. *Learning with computers: Analysing productive interaction.* Psychology Press.

Malmir, M., D. Forster, K. Youngstrom, L. Morrison and J. R. Movellan. 2013. Home Alone: Social Robots for Digital Ethnography of Toddler Behavior. In *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on.* IEEE pp. 762–768.

Marsella, S. C. and J. Gratch. 2009. "EMA: A process model of appraisal dynamics." *Cognitive Systems Research* 10(1):70 – 90. Modeling the Cognitive Antecedents and Consequences of Emotion.

Merrill, D. C., R. J. Reiser, M. Ranney and J. G. Trafton. 1992. "Effective Tutoring Techniques: A Comparison of Human Tutors and Intelligent Tutoring Systems." *The Journal of the Learning Sciences* 2(3):pp. 277–305.

Moundridou, M. and M. Virvou. 2002. Evaluating the Persona Effect of an Interface Agent in an Intelligent Tutoring System. In *Journal of computer assisted learning.* pp. 253–261.

Movellan, J., M. Eckhardt, M. Virnes and A. Rodriguez. 2009. "Sociable Robot Improves Toddler Vocabulary Skills." pp. 307–308.

Movellan, J. R., F. Tanaka, I. R. Fasel, C. Taylor, P. Ruvolo and M. Eckhardt. 2007. "The RUBI project: A Progress Report." *2nd ACM/IEEE International Conference on Human-Robot Interaction* pp. 333–339.

Nagao, T. and N. Ueda. 1996. NP-Completeness Results for NONOGRAM via Parsimonious Reductions. Technical report Tokyo Institute of Technology.

Nass, C., J. Steuer and E. R. Tauber. 1994. Computers Are Social Actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence.* pp. 72–78.

Nero, Shondel J. 2005. "Language, Identities, and ESL Pedagogy." *Language and Education* 19(3):194–211.

Nintendo. 2007*a*. "Picross DS." Nintendo DS Cartridge.

Nintendo. 2007*b*. "Wii Balance Board Peripheral." Nintendo Wii Fit.

Nkambou, R., J. Bourdeau and V. Psyché. 2010. "Building Intelligent Tutoring Systems: An Overview." *Advances in Intelligent Tutoring Systems* pp. 361–375.

Nwana, H. 1990. "Intelligent Tutoring Systems: An Overview." *Artificial Intelligence Review* 4(4):251–277.

Osgood, C. E., G. J. Suci and P. H. Tannenbaum. 1957. *The Measurement of Meaning.* Univ. of Illinois Press.

Pane, J. F., B. A. Griffin, D. F. McCaffrey and R. Karam. 2014. "Effectiveness of Cognitive Tutor Algebra I at Scale." *Educational Evaluation and Policy Analysis* 36(2):127–144.

Park, S. J., J. H. Han, B. H. Kang and K. C. Shin. 2011. Teaching assistant robot, ROBOSEM, in English class and practical issues for its diffusion. In *Advanced Robotics and its Social Impacts (ARSO), 2011 IEEE Workshop on.* pp. 8–11.

Prendinger, Helmut, Sonja Mayer, Junichiro Mori and Mitsuru Ishizuka. 2003. Persona Effect Revisited. In *Intelligent Virtual Agents*, ed. T. Rist, R. S. Aylett, D. Ballin and J. Rickel. Vol. 2792 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 283–291.

Prentzas, Jim. 2013. Artificial Intelligence Methods in Early Childhood Education. In *Artificial Intelligence, Evolutionary Computing and Metaheuristics*, ed. X. Yang. Vol. 427 of *Studies in Computational Intelligence* Springer Berlin Heidelberg pp. 169–199.

Pressley, M., E. Wood, V. E. Woloshyn, V. Martin, A. King and D. Menke. 1992. "Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning." *Educational Psychologist* 27(1):91–109.

Rus, V., S. K. D'Mello, X. Hu and A. C. Graesser. 2013. "Recent Advances in Conversational Intelligent Tutoring Systems." *AI magazine* 34(3):42–54.

Ruvolo, P., J. Whitehill, M. Virnes and J. Movellan. 2008. "Building a More Effective Teaching Robot Using Apprenticeship Learning." *7th IEEE International Conference on Development and Learning* pp. 209 –214.

San Pedro, M., R. Baker, S. Gowda and N. Heffernan. 2013. Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System. In *Artificial Intelligence in Education*, ed. H. C. Lane, K. Yacef, J. Mostow and P. Pavlik. Vol. 7926 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 41–50.

Saunders, G. 1988. *Bilingual Children: From Birth to Teens*. ERIC.

Schulze, K. G., R. N. Shelby, D. J. Treacy, M. C. Wintersgill, K. Vanlehn and A. Gertner. 2000. "Andes: An intelligent tutor for classical physics." *Journal of Electronic Publishing* 6(1).

Steele-Johnson, D. and B. G. Hyde. 1997. "Advanced technologies in training: Intelligent tutoring systems and virtual reality.".

Suebnukarn, S. and P. Haddawy. 2004. A Collaborative Intelligent Tutoring System for Medical Problem-based Learning. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*. IUI '04 New York, NY, USA: ACM pp. 14–21.

Sung, J., R. E. Grinter and H. I. Christensen. 2009. ""Pimp My Roomba": designing for personalization." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* pp. 193–196.

Tapus, A., C. Tapus and M. J. Matarić. 2009. "The role of physical embodiment of a therapist robot for individuals with cognitive impairments." *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on* pp. 103–107.

Thomaz, A. L. and C. Breazeal. 2008. "Teachable robots: Understanding human teaching behavior to build more effective robot learners." *Artificial Intelligence* 172(6):716–737.

Thomaz, A. L., G. Hoffman and C. Breazeal. 2006. Reinforcement Learning with Human Teachers: Understanding How People Want to Teach Robots. In *Proceedings the 15th IEEE International Symposium Robot and Human Interactive Communication (RO-MAN)*.

Topping, K. and S. Ehly. 1998. *Peer-assisted learning*. Routledge.

UGOBE. 2008. PleoWorld - The Home of Pleo, the Robotic Baby Dinosaur from UGOBE Life Forms. In *http://www.pleoworld.com*.

U.S. Bureau of Labor Statistics. 2014. "Unemployment Rate - Hispanic or Latino - LNS14000009." **http://data.bls.gov/cgi-bin/surveymost?ln**.

U.S. Census Bureau. 2011. "2010 Census." U.S. Department of Commerce.

Van Mulken, S., E. André and J. Müller. 1998. The persona effect: how substantial is it? In *People and computers XIII.* Springer pp. 53–66.

VanLehn, K. 2011. "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems." *Educational Psychologist* 46(4):197–221.

Wainer, J., D. J. Feil-Seifer, D. A. Shell and M. J. Matarić. 2007. "Embodiment and Human-Robot Interaction: A task-based perspective." *IEEE Proceedings of the International Workshop on Robot and Human Interactive Communication* pp. 872–877.

Welch, L. R. 2003. "Hidden Markov models and the Baum-Welch algorithm." *IEEE Information Theory Society Newsletter* 53(4):10–13.

Williams, S. M., D. Nix and P. Fairweather. 2013. Using speech recognition technology to enhance literacy instruction for emerging readers. In *Fourth International Conference of the Learning Sciences.* pp. 115–120.

Wood, D., J. S. Bruner and G. Ross. 1976. "The role of tutoring in problem solving." *Journal of child psychology and psychiatry* 17(2):89–100.

Yun, Sangseok, Jongju Shin, Daijin Kim, ChangGu Kim, Munsang Kim and Mun-Taek Choi. 2011. Engkey: Tele-education Robot. In *Social Robotics*, ed. B. Mutlu, C. Bartneck, J. Ham, V. Evers and T. Kanda. Vol. 7072 of *Lecture Notes in Computer Science* Springer Berlin Heidelberg pp. 142–152.

Zeng, Z., M. Pantic, G. I. Roisman and T. S. Huang. 2009. "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(1):39–58.