The Effect of Personalization in Longer-Term Robot Tutoring

DANIEL LEYZBERG, Princeton University, USA
ADITI RAMACHANDRAN and BRIAN SCASSELLATI, Yale University, USA

The benefits of personalized social robots must be evaluated in real-world educational contexts over periods of time longer than a single session to understand their full potential to impact learning outcomes. In this work, we describe a personalization system designed for longer-term personalization that orders curriculum based on an adaptive Hidden Markov Model (HMM) that evaluates students' skill proficiencies. We present a study investigating the effectiveness of this system in a five-session interaction with a robot tutor, taking place over the course of 2 weeks. Our system is evaluated in the context of native Spanish-speaking first-graders interacting with a social robot tutor while completing an English Language Learning educational task. Participants either received lessons: (1) ordered by our adaptive HMM personalization system which selects a lesson based on a skill that the individual participant needs more practice with ("personalized condition") or (2) ordered randomly from among the lessons 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.

CCS Concepts: • **Applied computing** → **Interactive learning environments**; • **Human-centered computing** → *User models*; *User studies*;

Additional Key Words and Phrases: Human-robot interaction, personalization, tutoring, English Language Learning (ELL)

ACM Reference format:
Daniel Leyzberg, Aditi Ramachandran, and Brian Scassellati. 2018. The Effect of Personalization in Longer-Term Robot Tutoring. *ACM Trans. Hum.-Robot Interact.* 7, 3, Article 19 (December 2018), 19 pages.
https://doi.org/10.1145/3283453

1 INTRODUCTION

Robot tutoring systems have demonstrated great potential as effective tutoring agents in the educational domain [5]. Recent work indicates that physically embodied tutoring agents can increase cognitive learning gains, as well as boost enjoyment and compliance [2, 37, 43]. The social presence of a robot tutor can also put students at ease during learning, positively impacting students' social behaviors [22]. Another promising capability of robot tutoring systems is providing personalization for an individual within an interaction, which can also strengthen learning outcomes. Relatively straightforward personalization strategies for providing lessons have been shown to

Authors' addresses: D. Leyzberg, Princeton University, 35 Olden St, Princeton, NJ 08544; email: dan.leyzberg@princeton.edu; A. Ramachandran and B. Scassellati, Yale University, 51 Prospect St, New Haven, CT 06511; emails: {aditi.ramachandran, brian.scassellati}@yale.edu.

2573-9522/2018/12-ART19
https://doi.org/10.1145/3283453

positively impact learning gains during a cognitive puzzle-solving task [36]. Though this body of research indicates the clear benefits of personalized robot tutors, many of these studies deal with adults, lab-based interactions, or single-session interaction contexts.

To more thoroughly understand how personalized robot tutors can impact learning gains, it is crucial that we design and study longer-term interactions in real-world learning contexts. Furthermore, it is important to focus on educational problems with high-demand to accurately assess the need for these systems—for example, to assist young children who need the additional educational help. The role of personalized social robot tutors within these diverse learning contexts needs to be more thoroughly explored. To this end, we seek to gain a better understanding of the effectiveness of robot tutoring systems in high-impact application domains over periods of time longer than a single session. We want to evaluate whether the observed learning benefits of personalization of content within short-term interactions are also observed for longer-term interactions involving children.

In this work, we investigate the impact of personalized robot tutoring specifically on learning outcomes in a long-term educational interaction. We designed a personalization system that uses an adaptive Hidden Markov Model (HMM) to order content within an interactive English Language Learning (ELL) story task based on the skill proficiencies of students. We detailed our interaction context design, which includes an interactive story with interchangeable chapters that requires students to verbally complete Spanish-to-English translation exercises. We conducted a long-term human–robot interaction study to evaluate the effectiveness of our personalization system over the course of five tutoring sessions by comparing a personalized condition to a non-personalized one. Our results show that children who received personalized lessons during the robot tutoring sessions significantly outperformed those who received non-personalized lessons and that our personalization system successfully identified student skills requiring practice.

2 BACKGROUND

In this section, we provide the relevant background on ELL education and why we chose to study our personalized tutoring system in the context of this educational domain. We also review related work from both the intelligent tutoring systems (ITS) community as well as the human–robot interaction (HRI) community involving personalized learning systems.

2.1 ELL Education

According to the 2010 United States Census data, 20% of American households speak a language other than English in the home [55]. Children raised in non-native English-speaking households can face a preparatory disadvantage in school relative to their native-speaking peers and research in education has frequently referenced an achievement gap between students proficient in English and those that are English learners [47, 54]. 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 [7].

Effective ELL education is vital to leveling the playing field for children raised in non-native English speaking homes. Though there are many successful programs supplying ELL education across the country, especially in major metro areas like New York and Los Angeles, millions of students still receive little or poor-quality ELL education [23]. Due to the necessity of effective ELL instruction, in this study, we focus on teaching an English Language Learning task to children in first grade (typically around ages 5 to 6). To evaluate the effectiveness of a personalized tutoring system, we wanted to focus on an authentic teaching task with children who may benefit from this type of instruction. 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 [24].

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 [24]. We chose to work with first-grade students for this reason.

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 ELL instruction to Spanish-dominant first-grade students in a bilingual elementary school.

2.2 Intelligent Tutoring Systems

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 ITS community. For this work, we made a curriculum-sequencing tutor that does not provide step-by-step feedback. Instead, this tutor sequences an individualized path through available curriculum to maximize the effectiveness of the lessons for each student [14]. Though various tutoring systems from the ITS community have explored the impact of curriculum-sequencing on learning outcomes, our work focuses on understanding the benefits of personalized curriculum-sequencing embedded within an interactive robot tutoring ELL task.

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 [40]. 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 [10, 15] to detecting if students are trying to abuse the hint and help features to game the system to improve their scores [3]. The most significant type of personalization in automated tutors is the student model [21]. The student model used in this article extends the Bayesian Knowledge Tracing (BKT) family of student models originated by Corbett and Anderson for use in a personalized robot tutoring system designed for the ELL domain [12]. Many ITSs utilize models derived from BKTs that typically model a student's knowledge state during the skill acquisition, where each individual skill is considered to be known or unknown to the student at each opportunity to practice a given skill [12]. Based on observations of the student's performance on a given skill, these probabilities are updated to reflect which skills the student may not know or need extra practice on. This is the most commonly used technique to model student knowledge within automated tutoring systems and has been successfully used in a variety of tutoring applications including tutoring for math and programming skills [12, 29]. There are also more complex models that can also be used to model a student's memory to plan practice opportunities accordingly for given skills [38]. Rather than represent individual skills as being either known or unknown, other models such as activation-based memory models typically represent learning as individual skills that are acquired gradually over time where each item is remembered based on the frequency and recency of practice opportunities and forgotten as a function of time [41]. Some of these models have been shown to be successful in language learning tasks and have typically been used to space out practice opportunities for vocabulary memorization tasks [41, 42].

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 overview see Reference [34]. 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 [35]. A common paradigm in CALL research is systems that process the speech of the user and correct errors in pronunciation, prosody, or grammar [16]. 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 [39].

2.3 Personalized Robot Tutoring

Recent research has demonstrated the effectiveness of social robots as educational agents that foster learning benefits due to personalized interaction [20]. Social robots can personalize one-on-one tutoring interactions for students based on a variety of dimensions, including a student's nonverbal behavior, learning style, and task performance [9]. Adults receiving personalized lessons from a robot tutor in a single session significantly improved their performance on a cognitive puzzle task as compared to those who received non-personalized lessons [36]. The personalized lessons were chosen based on a Bayesian network skill assessment model, which is closely related to the HMM model we use in this work. Social robot tutors have also been used to accurately assess children's word-reading skills based on a Bayesian active-learning algorithm, and personalize aspects of a story-telling interaction with young children [18]. Other studies have investigated the personalization of aspects of a learning interaction that are not directly related to the content, such as the social behavior of the robot tutor. For example, children interacting with a social robot with higher nonverbal immediacy and responsiveness demonstrated greater learning gains when compared to a less immediate robot tutor [27]. Children also exhibited significant learning gains when given breaks based on personalized timing strategies as compared to a fixed timing strategy during a cognitively taxing learning interaction [45]. Though these studies each explore interesting aspects of personalization within a learning interaction, they all measure the benefits of personalization over the course of a single tutoring session. Our work seeks to isolate the benefits of personalization within a robot tutoring interaction over a longer-term interaction.

There has been some work investigating the personalization within robot tutoring interactions in a longer-term setting. A robot learning companion that displayed empathic behaviors towards the user over a long-term interaction positively impacted perceptions of the robot over time [33]. Students who interacted with an adaptive robot tutor that regulated their help-seeking behavior during a math tutoring interaction showed improved behavior and learning outcomes over multiple tutoring sessions [46]. A social robot that personalized its affective reactions to each child over the course of two months was shown to impact student valence as compared to a non-personalizing robot in a language-learning context [19]. Several studies involving personalized robots in learning have also explored methods to sustain engagement for students, for example, by leveraging attention data from wearable sensors [50] or by employing multi-activity switching over the course of several interactions [11]. More recent work has gone farther by investigating a robot tutor that provides personalized feedback and encourages students to build self-regulated learning skills over multiple sessions, demonstrating that students who interacted with the personalized robot that provided scaffolding were able to more successfully enhance their self-regulated learning skills [25]. These studies address the effects of various supportive mechanisms within a learning interaction, whereas our work aims to understand the effect of content-based personalization with a robot tutor over multiple sessions on learning outcomes. Furthermore, a more comprehensive long-term field study demonstrated that students showed higher learning gains when interacting with a robot with multiple dimensions of personalization including its nonverbal behavior, its verbal behavior, and its content progression as compared to students interacting with a non-personalized version of the same robot [4]. While one of the personalization components in this study did involve content delivery, it utilized a fixed threshold to determine mastery of a given topic, as compared to our work which utilizes a personalized model of the child's knowledge of the given skill to provide practice opportunities accordingly.

In recent years, there have also been advances in robot tutoring systems that focus specifically on language-learning tasks [6]. Social robots have been shown to foster lasting learning gains in a second-language tutoring task when compared to a no-robot baseline [28]. Students learning English as a foreign language retained more vocabulary words when their traditional instruction was supplemented with a robot tutor than when they did not receive additional time with the robot over a long-term interaction of five sessions [1]. Children who practiced English language skills with robot tutors designed to help language skills improved their speaking skills, and these students reported increased interest, confidence, and motivation in learning English after spending multiple tutoring sessions with the robot [31]. Specifically regarding content personalization techniques, new research exploring a content personalization approach based on an extension of Bayesian Knowledge Tracing was shown to help adults learn new words in an artificial language when compared to a random baseline for content selection [48]. These personalization techniques were also applied to children in a single-session language tutoring task where children completed a vocabulary exercise by identifying the correct animal on a screen during an interaction with a robot. Results from this study indicated that children who received content based on the adaptive content selection algorithm and a robot that displayed gestures demonstrated higher learning performance during the interaction; however, these learning gains were not observed independent of the interaction when measured from pre-test to post-test [13]. Building on the related work demonstrating the efficacy of robot tutors in language learning scenarios, we specifically investigate the effects of a robot tutor that personalizes its content progression in a real-world language learning task for children over multiple tutoring interactions.

3 METHODOLOGY

In this section, 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 in first grade. 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 evaluated students before they participated in this study 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.

3.1 Interaction Content

During the course of this experiment, the robot engaged participants in an interactive adventure story task. 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. An example translation task can be seen in Figure 1. All of the translation tasks participants did in this study were sentences that, in English, contain either the words "make" or "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 cake today" instead of "I made a cake today."

In the English language, there are as many as 10 distinct categories of usage for these two words that distinguish them from one another, depending on the ELL curriculum one chooses. For this

Fig. 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 either 30 or 40 translations per session, with sessions lasting approximately 20 minutes. Each participant did five sessions over the course of two weeks.

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 1. We chose to teach 4 categories rather than teaching all 10 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 3.5. 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, then the translation was marked "incorrect." If neither verb was used in the translation, then it was marked "irrelevant." If the participant did not respond, then "silent" was marked. No explicit feedback was given after each answer by the either the experimenter or the robot. Rather, each translation exercise was an opportunity to practice a given skill.

3.2 Experimental Procedure

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

The experimental procedure was designed to meet ethical guidelines and was approved by an Institutional Review Board before it was conducted. Before the experiment began, a voluntary consent form was sent to parents of potential participants, all of whom were in the same

Table 1. The English Words "Make" and "Do" Translate to One Word in Spanish ("hacer"), and as a Result Many Native Spanish Speakers Struggle to Learn the Distinction We Make between These Words When They Learn English

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

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.

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 study by the experimenter. Participants engaged in five sessions of approximately 20 minutes in length, no more than once per day, over the course of 2 weeks. 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 fourth 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 4 grammatical skills described above. We authored a total of 24 interchangeable chapters for this study, 6 that targeted each of the 4 skills. Each of the chapters were designed to include translation tasks of the same general difficulty level. In total, each participant saw only 9 of the 24 interchangeable chapters. This limitation was necessitated by the population to avoid fatigue. For a visual representation of the content of each session see Figure 2. The ordering and selection of the lessons was determined by the condition the participant was in.
- The fifth and last session of the study was a post-test. Like the pre-test, there were 40 translations, 10 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 5.

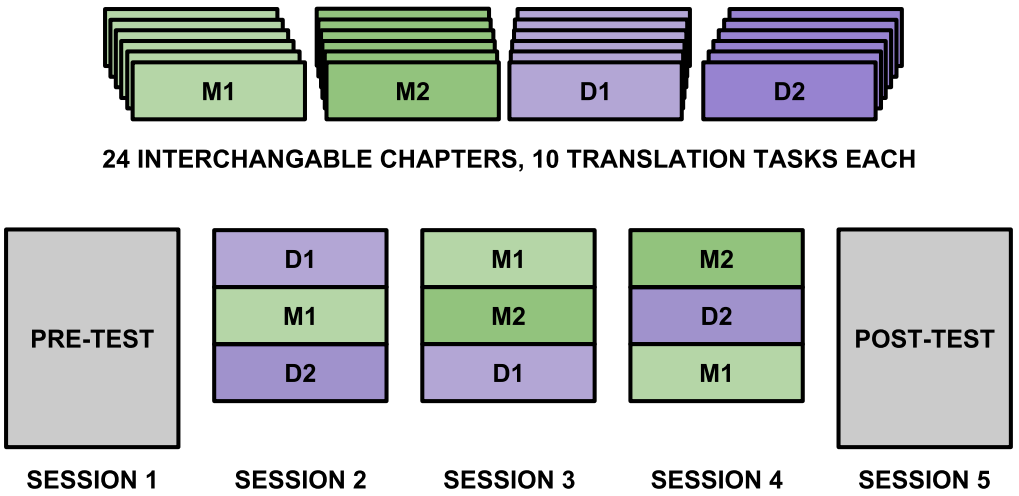

Fig. 2. Participants engaged in five sessions over the course of a 2-week period. The first session was a pre-test, the same across all participants, with 10 translation tasks per skill. The middle three sessions were composed of three interchangeable chapters, each focused on one specific skill, and 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 in Section 4 below.

3.3 Experimental Design

We designed a between-subjects study to understand the benefits of our personalized tutoring system. The only independent variable in this study was the ordering of the interchangeable lesson chapters during sessions two through four. Below, we provide descriptions of each of our two experimental conditions: *personalized* and *non-personalized*.

3.3.1 Personalized Condition. In the *personalized* lessons condition, the episodes were ordered based on a HMM that we built for each participant and skill. The model for each skill consisted of three hidden states, (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 4 below. If no skills were "not-known," then lessons that targeted "forgotten" skills were chosen randomly among those not yet seen. Last, if all of the skills were "known," then the tutor chose a random episode among the ones the participant had not yet seen.

3.3.2 Non-Personalized Condition. In the *non-personalized* lessons condition, participants received a random episode that they had not yet seen, distributed uniformly over the four skills. Because participants saw nine 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.

3.4 Robot

The robot we used for this study, 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

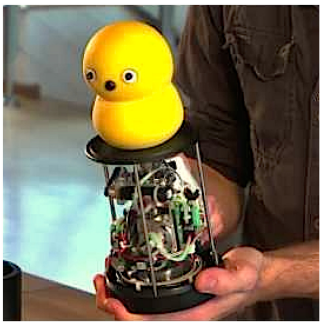

Fig. 3. A Keepon robot. Keepon can rotate left and right, lean side to side, tilt up and down, and bounce up and down. In this study, Keepon interacts one-on-one with a student in a tutoring context.

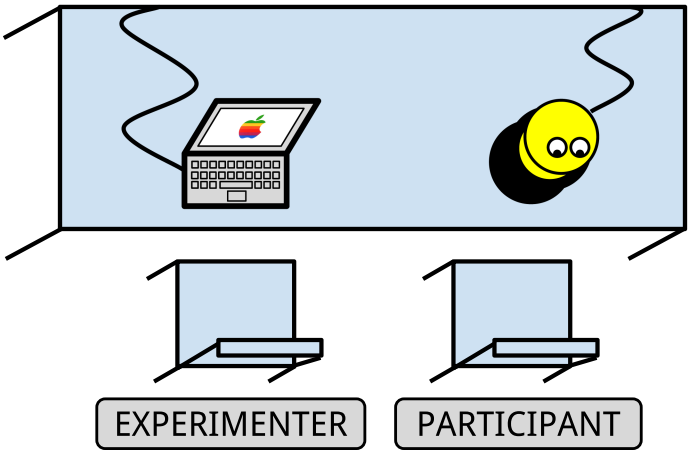

Fig. 4. Overhead view of the experimental apparatus. The participant, a first-grade student whose dominant language is Spanish, is seated facing the robot. The experimenter, who provides adult supervision and natural language processing for the robot, is seated beside the participant. The experimenter provided occasional vocabulary assistance, as well as categorizing each of the participant's responses as correct, incorrect, irrelevant, or silent.

a microphone. For a photograph, see Figure 3. In this study, the robot faced the participant and bounced while speaking in a personalized ordering of pre-recorded Spanish audio clips. See Figure 1 for the relative positioning of the robot and the participant.

3.5 Experimenter

The participant was in the constant supervision of an adult during the course of this study. This adult, the first author, also played a role in the experiment. The experimenter and the participant sat side by side as seen in Figure 4. The experimenter performed three roles as follows:

(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 systems to process the participants' speech, because such systems have relatively high error rates with children and non-native

speakers [8, 58]. Instead, the experimenter provided speech recognition information to the system by coding each of the participants' responses as "correct," "incorrect," "irrelevant," or "silent" using the objective rules described in Section 3.1.

(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, to preserve the integrity of the "make" vs. "do" distinction made entirely by participants.

3.6 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 in the same first-grade class at the same school. The participants were exclusively Spanish-dominant speakers, being raised in Spanish-dominant homes. We verified with the teacher of these students that study content (translation exercises within a story format) was at an appropriate level for all students who participated. Additionally, the teacher categorized the students as being a homogeneous group in terms of their individual ELL skill levels. Participants were randomly divided into the two experimental conditions.

4 PERSONALIZATION MODEL

There were two conditions in this study, personalized lessons and non-personalized lessons. Below we provide a detailed account of the model used to provide the personalized ordering of lessons for the *personalized* condition.

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 3.5, 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. Our modeling approach is based on the most commonly used student modeling technique in ITS research, a family of models called BKT [12]. Though many other models exist, some of which offer ways to account for more complex aspects of knowledge acquisition and memory [38, 41, 42], this simpler type of modeling technique has not been thoroughly investigated in the context of a longer-term robot tutoring interaction for children in a real-world educational setting. We deliberately designed our personalization model to be derived from the family of BKT models to understand whether this type of personalization model could impact learning outcomes in the context of an ELL task for children.

For each skill and each participant, we created independent same-structured 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 5. There were four observable states in this model: (1) a correct answer, (2) incorrect answer, (3) irrelevant answer, or (4) no answer. 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 [57]. 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

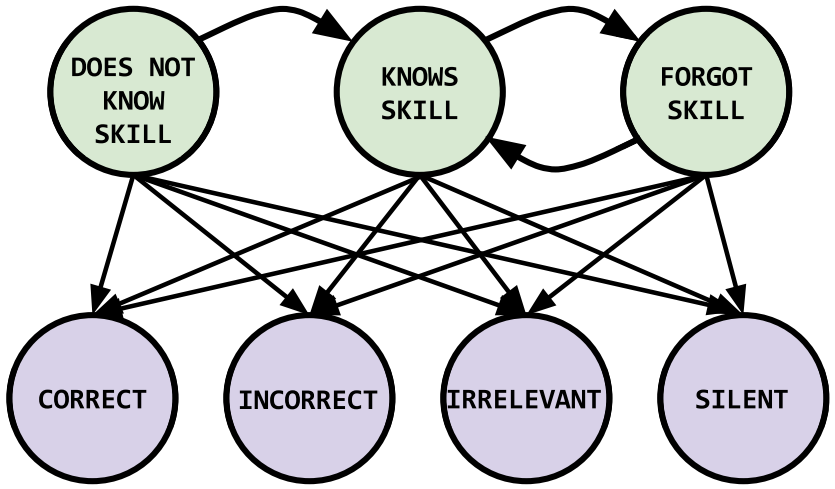

Fig. 5. The HMM used to sequence curriculum for the personalized group. Four simultaneous copies of this model were trained and run for each student, one for each of the English grammar skills defined above. Implementation details of the HMM can be found in Section 4.

states for all four skills, based on the expert estimate of an ELL 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 observation 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 [30]. 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 $P(CORRECT|KNOWS) = 0.9$ for all four skills.

We apply the Viterbi algorithm which gives a state for each of the observations resulting in the most likely sequence of 10 states [17]. We use the latest state to pick the most likely hidden state for the given skill. This tells us which of the four skills each student knows, does not know, or has forgotten, and we use that information to choose a personalized lesson for each participant in the following order:

- If any skill is unknown, then 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, then 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 a result of our experience running this experiment with a pilot group over the course of 5 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 5 below.

5 RESULTS

We investigated the effects of our personalization system on a longer-term robot tutoring interaction. Participants performed either 30 or 40 translation tasks per session and an experimenter coded each translation as "correct", "incorrect", "irrelevant", or "silent." Each student is assigned an accuracy score for each session to measure performance on the translation tasks. This score is defined as the number of correct translations divided by the total number of translations in a given session. We compare mean accuracy scores between groups to evaluate the impact of our personalization system. We used Student's t-tests for all comparisons of performance between groups and used an α level of .05 to determine significance for all statistical tests reported.

5.1 Personalization Impacts Learning Gains

To evaluate the performance of the two experimental groups over the course of the multiple tutoring sessions, we first compared the pre-test scores between groups and then compared the post-test scores of both groups. There was no significant difference in the pre-test scores between the two experimental groups, with mean scores of ($M = .38, SD = .11$) for the non-personalized group and ($M = .36, SD = .13$) for the personalized group. Participants who received personalized lessons ($n = 10$) performed significantly better on the post-test ($M = .84, SD = .08$) than participants who received non-personalized lessons (n=9) ($M = .63, SD = .09$), $t(17) = 2.368$, $p = .030$, $d = 2.47$ (Figure 6). These results indicate 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. Furthermore, 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.

Another result of our personalization system is the difference 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 non-personalized group. These data are plotted in Figures 7 and 8. The mean accuracy score was significantly lower in the personalized lessons group, ($M = .28, SD = .08$), than in the non-personalized group, ($M = .50, SD = .10$), $t(17) = -2.898$, $p = .010$, $d = 2.43$. 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 to challenge students.

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 7, where the personalized participants response distributions are plotted, there is a steep growth in the amount of "correct" answers from the second session to the fifth session. 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

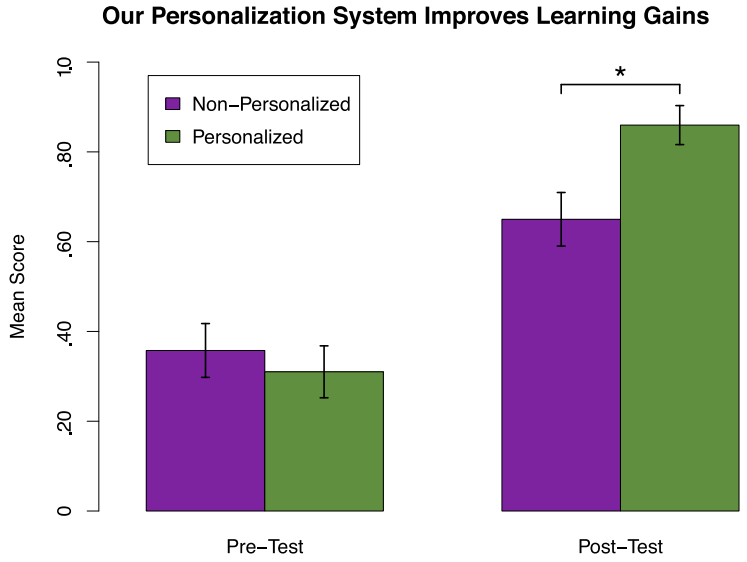

Fig. 6. Pre-test and post-test results across experimental groups, indicating the effectiveness of our personalization system. Participants who received personalized lessons performed significantly better on the post-test than participants who received non-personalized lessons. (*) denotes $p < .05$. Error bars depict standard error.

thus failed more often early in the study, their post-test scores were very high. In Figure 8, which shows the corresponding data for the participants who received non-personalized lessons, we see a growth pattern that steadily increases but less drastically as compared to the personalized group. 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.

5.2 Additional Observations

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 10 of 40 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 ELL domain. When personalization is added to a robot language tutor, which would be useful on its own, the gains are even higher.

Another interesting outcome of this study is the number of "incorrect" answers among the data in either group. The mean overall occurrence of "incorrect" answers across both groups and all sessions was 7% ($SD = 3\%$). For a visual representation of the distribution of these answers per session and group, see Figures 7 and 8. 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. Though somewhat infrequent, students still made mistakes by giving incorrect responses when interacting with the robot. Young students who are English learners often feel anxiety when interacting with

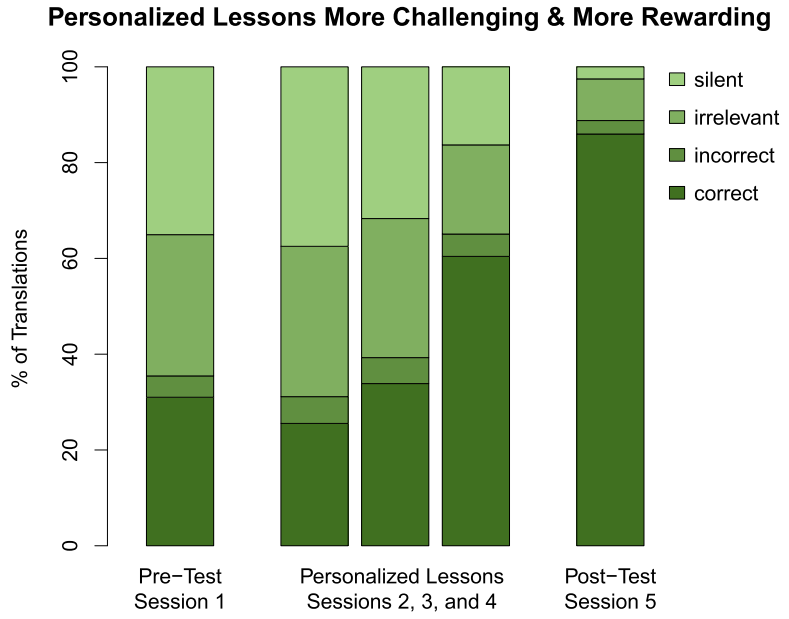

Fig. 7. Distribution of answers given by participants in the personalized lessons condition across all five sessions with the robot. Between the pre-test and post-test, we see a momentary drop in accuracy scores that then seem to rapidly increase until the post-test. This result validates our main manipulation, 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.

other native English speakers due to various internal and external pressures [52]. It is possible that students may perceive the robot differently from that of another person and could be more willing to make mistakes in the presence of the robot. This idea is related to prior work that investigated the roles of assistant versus teacher involving both robots and humans and how this impacts student help-seeking behavior, and concluded that students may have felt less comfortable asking for help from a human teacher as compared to the human assistant, the robot assistant, and the robot teacher [22]. However, other work has shown that students ask for help more often from a human tutor than from a robot tutor, indicating the need to understand student perceptions of the robot's capabilities prior to an interaction [49]. Toward the end of the study, students made fewer mistakes and reduced the number of times they said something irrelevant and stayed silent, likely because participants had more knowledge of the skills.

6 DISCUSSION

The success of our personalization approach of choosing the sequence of chapters to present to each student contributes to the growing body of research in the HRI community highlighting the importance of personalization mechanisms with educational interactions. Earlier work showed this type of approach could be used to trace student knowledge during a puzzle task and provide hints accordingly resulting in faster performance for adults [36]. More recent work has looked at an adaptive approach to content selection also based on BKT models and has shown that adults perform better with this approach applied specifically within a language learning setting [48]. Further research using the same model was applied to a real language learning scenario with

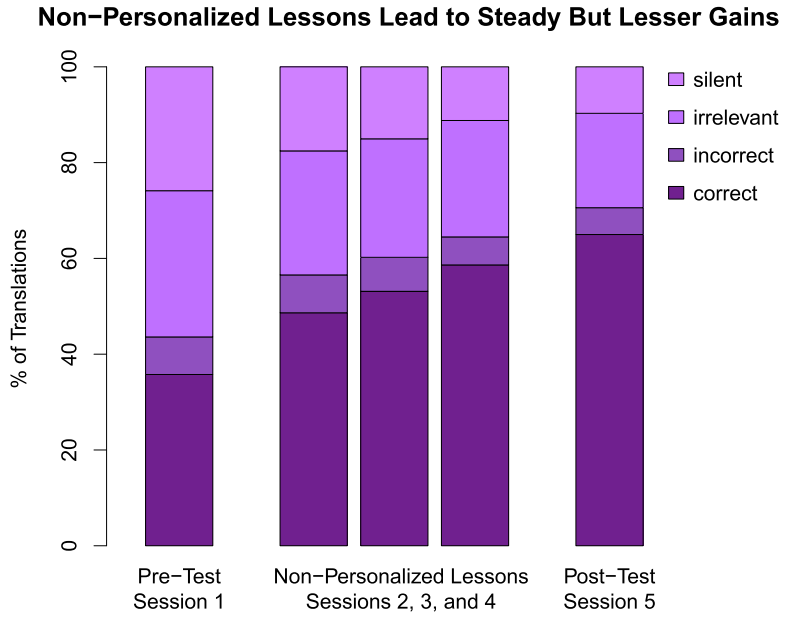

Fig. 8. Distribution of answers given by participants in the non-personalized lessons condition across all five sessions with the robot. Between the pre-test and post-test, we see a steady increase in correct responses, and corresponding decrease in the incorrect, irrelevant, and silent responses. This 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.

children and it was successful, particularly when gestures were combined with the adaptive content selection approach [13]. The results from our study further validate these related results and provide additional evidence towards the feasibility of a relatively simple personalization approach to content sequencing over multiple sessions. Our results not only validate the effectiveness of a content personalization model in a language learning setting, but we apply this to children in an authentic ELL tutoring task and evaluate the model's impact on learning outcomes over multiple tutoring sessions rather than a single session. Based on our model design, we would expect a more complex model of knowledge and memory that informs the personalization approach to yield similar or better results. It is promising that a personalized content sequencing approach based on a simpler family of models can still promote strong learning gains.

Students who participated in our study stayed engaged with the robot throughout each of the five sessions and enjoyed their time with the robot as we observed that they smiled and laughed during the sessions. Though this evidence is anecdotal, it is in line with what other studies that have explored longer-term robot interactions with children have found: Children typically stay engaged with robots over multiple sessions and often form social bonds with the robot [26, 32, 51]. Other work investigating on multiple-session robot interactions with children found that multi-activity switching can be used to maintain engagement over multiple interactions [11]. Our work highlights the idea that simple content personalization over multiple tutoring sessions may help keep children engaged by providing them with exercises of the right level of challenge. This combined with other recent findings provides several design recommendations and approaches for robots to successfully engage with children over time, improving learning outcomes. Despite these advances, additional research investigating truly long-term interactions, involving months

or years of interaction time designed in conjunction with qualified educational authorities is necessary to understand whether we can successfully deploy autonomous robot tutors in classrooms and homes.

In addition, we must interpret our results carefully, making sure not to over-attribute the success of the learning outcomes to the robot tutoring platform specifically. Our analysis and experimental design showcases the value of a personalized approach to sequence curriculum within a ELL tutoring task and specifically highlights the advantages of this model over utilizing a robot tutor with a non-personalized content selection method. It was not our focus in this study to understand the value of the presence of the robot tutor itself, as we relied on prior work [37, 44, 56] that indicates the value of physically present robots. However, one limitation of the study is the fact that we did not have the resources to also conduct a no-robot baseline control condition to isolate both the effects of the robot tutor as well as the personalization approach, which other work in this area has done [28]. Nonetheless, the no-robot baseline in the study conducted by Kennedy et al. did not compare the same tutoring provided through another medium (such as a tablet or a computer screen) and the robot tutoring interaction. This demonstrates the value of the robot tutoring interaction as a whole, but not specifically the robot. Though we do have evidence from multiple sources about the benefits physically embodied agents [2, 37, 43, 53], more work should still be done to validate these findings and further explore more precisely when a robot is necessary as well as what about the robots lead to more effective learning in certain contexts.

We also note that our personalization model performed better than a ”random” baseline. However, in our case, our control condition is not truly random, as it it does not present content in an order that does not make sense. For curriculum sequencing approaches in particular, the control condition can be well represented by what teachers do in classrooms when they must take a one-size-fits-all approach. Similarly, in our user study, we designed a control condition that emulates how a teacher generally presents content to a class without ordering it based on an individual's knowledge of particular skills. Though we did use a “random” baseline, we feel that in this context it was a strong control condition, thereby highlighting the value of the personalization model that corresponded to a larger learning improvement than the control.

Another challenge in designing robot tutors for these type of educational environments involving children is handling individual differences. The students who participated in this study were quite homogeneous in terms of skill level on the translation task that was involved in the study. We used our personalization model to track student knowledge on the given skills and present content accordingly. However, learning is a complex process that often involves several other salient factors that can contribute to student performance and efficacy in learning, such as affective factors, attention, and personality traits. Other work involving various applications of human–robot interaction has investigated some of these aspects that are relevant in a tutoring setting for children [19, 33, 50]. Studying each of these aspects individually is important; however, to build effective, personalized tutors that can be used over longer term interactions, future work should investigate personalization mechanisms that take many of these important factors into account, rather than just one.

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 ELL 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 extent, to the students' lives. As such systems become more robust, future work should include follow-up research to see what the long-term impacts are of such interventions.

7 CONCLUSION

In this work, we describe a personalization system for longer-term robot tutoring and we test our system with ELL 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 ELL 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 student's weakest skills and can be used to sequence curriculum to maximize a robot tutor's effectiveness.

REFERENCES

[1] Minoo Alemi, Ali Meghdari, and Maryam Ghazisaedy. 2014. Employing humanoid robots for teaching English language in Iranian junior high-schools. *Int. J. Human. Robot.* 11, 3 (2014). DOI: https://doi.org/10.1142/S0219843614500224

[2] Wilma A. Bainbridge, Justin Hart, Elizabeth S. Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'08)*. IEEE, 701–706.

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

[4] Paul Baxter, Emily Ashurst, Robin Read, James Kennedy, and Tony Belpaeme. 2017. Robot education peers in a situated primary school study: Personalisation promotes child learning. *PLOS ONE* 12, 5 (2017), e0178126. DOI: https://doi.org/10.1371/journal.pone.0178126

[5] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Sci. Robot.* 3, 21 (2018). http://robotics.sciencemag.org/content/3/21/eaat5954.

[6] Tony Belpaeme, Paul Vogt, Rianne van den Berghe, Kirsten Bergmann, Tilbe Göksun, Mirjam de Haas, Junko Kanero, James Kennedy, Aylin C Küntay, Ora Oudgenoeg-Paz, et al. 2018. Guidelines for designing social robots as second language tutors. *Int. J. Soc. Robot.* 10, 3 (2018), 325–341.

[7] R. M. Callahan. 2005. Tracking and high school English learners: Limiting opportunity to learn. *Am. Educ. Res. J.* 42, 2 (2005), 305–328.

[8] M. Chen 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)*. Association for Computational Linguistics, Stroudsburg, PA, 722–731. http://dl.acm.org/citation.cfm?id=2002472.2002564

[9] Caitlyn Clabaugh, Gisele Ragusa, Fei Sha, and Maja Matarić. 2015. Designing a socially assistive robot for personalized number concepts learning in preschool children. In *Proceedings of the 2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob'15)*. IEEE, 314–319.

[10] C. Conati and H. Maclaren. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Model. User-Adapt. Interact.* 19, 3 (2009), 267–303.

[11] Alexandre Coninx, Paul Baxter, Elletra Oleari, Sara Bellini, Bert Bierman, Olivier Blanson-Henkemans, Lola Canamero, Piero Cosi, Valentin Enescu, Raquel Ros, Antoine Hiolle, Remi Humbert, Bernd Kiefer, Ivana Kruijff-Korbayova, Rosemarijn Looije, Marco Mosconi, Mark Neerincx, Giulio Paci, Georgios Patsis, Clara Pozzi, Francesca Sacchitelli, Hichem Sahli, Alberto Sanna, Giacomo Sommavilla, Fabio Tesser, Yiannis Demiris, and Tony Belpaeme. 2016. Towards long-term social child-robot interaction: Using multi-activity switching to engage young users. *J. Hum.-Robot Interact.* 5, 1 (2016), 32–67.

[12] Albert T. Corbett and John R. Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.* 4, 4 (1994), 253–278.

[13] Jan de Wit, Thorsten Schodde, Bram Willemsen, Kirsten Bergmann, Mirjam de Haas, Stefan Kopp, Emiel Krahmer, and Paul Vogt. 2018. The effect of a robot's gestures and adaptive tutoring on children's acquisition of second language vocabularies. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 50–58.

[14] M. C. Desmarais and R. S. Baker. 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Model. User-Adapt.Interact.* 22, 1–2 (Apr. 2012), 9–38. DOI: https://doi.org/10.1007/s11257-011-9106-8
[15] S. D'Mello. 2012. Monitoring affective trajectories during complex learning. In *Encyclopedia of the Sciences of Learning*. Springer, 2325–2328.
[16] M. Eskenazi. 2009. An overview of spoken language technology for education. *Speech Commun.* 51, 10 (2009), 832–844.
[17] G. D. Forney Jr. 1973. The Viterbi algorithm. *Proc. IEEE* 61, 3 (1973), 268–278.
[18] Goren Gordon and Cynthia Breazeal. 2015. Bayesian active learning-based robot tutor for children's word-reading skills. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'15)*. 1343–1349.
[19] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. 2016. Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'16)*. 3951–3957.
[20] Jillian Greczek, Elaine Short, Caitlyn E. Clabaugh, Katelyn Swift-Spong, and Maja Mataric. 2014. Socially assistive robotics for personalized education for children. In *Proceedings of the AAAI Fall Symposium on Artificial Intelligence and Human-Robot Interaction (AI-HRI'14)*.
[21] K. E. Hogan and M. E. Pressley. 1997. *Scaffolding Student Learning: Instructional Approaches and Issues*. Brookline Books.
[22] Iris Howley, Takayuki Kanda, Kotaro Hayashi, and Carolyn Rosé. 2014. Effects of social presence and social role on help-seeking and learning. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 415–422.
[23] K. Humes, N. A. Jones, and R. R. Ramirez. 2011. *Overview of Race and Hispanic Origin, 2010*. U.S. Department of Commerce.
[24] J. S. Johnson 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. *Cogn. Psychol.* 21, 1 (1989), 60–99.
[25] Aidan Jones and Ginevra Castellano. 2018. Adaptive robotic tutors that support self-regulated learning: A longer-term investigation with primary school children. *Int. J. Soc. Robot.* 10, 3 (2018), 357–370.
[26] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Hum.-Comput. Interact.* 19, 1 (2004), 61–84.
[27] James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme. 2015. Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions. In *Proceedings of the International Conference on Social Robotics*. Springer, 327–336.
[28] James Kennedy, Paul Baxter, Emmanuel Senft, and Tony Belpaeme. 2016. Social robot tutoring for child second language learning. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. IEEE, 231–238.
[29] Kenneth R. Koedinger. 2002. Toward evidence for instructional design principles: Examples from cognitive tutor math 6. In *Proceedings of the Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. ERIC, 21–49.
[30] C. C. Kulik, J. A. Kulik, and R. L. Bangert-Drowns. 1990. Effectiveness of mastery learning programs: A meta-analysis. *Rev. Educ. Res.* 60, 2 (1990), 265–299.
[31] Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, and Munsang Kim. 2011. On the effectiveness of robot-assisted language learning. *ReCALL* 23, 1 (2011), 25–58.
[32] Iolanda Leite, Ginevra Castellano, Andre Pereira, Carlos Martinho, and Ana Paiva. 2012. Long-term interactions with empathic robots: Evaluating perceived support in children. In *International Conference on Social Robotics*. Springer, 298–307.
[33] Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. 2014. Empathic robots for long-term interaction. *Int. J. Soc. Robot.* 6, 3 (2014), 329–341.
[34] M. Levy. 1997. *Computer-Assisted Language Learning: Context and Conceptualization*. ERIC.
[35] M. Levy and G. Stockwell. 2013. *CALL Dimensions: Options and Issues in Computer-assisted Language Learning*. Routledge.
[36] Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. 2014. Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 423–430.
[37] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati. 2012. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. Cognitive Science Society, Austin, TX.
[38] Robert V. Lindsey, Jeffery D. Shroyer, Harold Pashler, and Michael C. Mozer. 2014. Improving students' long-term knowledge retention through personalized review. *Psychol. Sci.* 25, 3 (2014), 639–647.
[39] K. Littleton and P. Light. 1999. *Learning with Computers: Analysing Productive Interaction*. Psychology Press.

[40] D. C. Merrill, R. J. Reiser, M. Ranney, and J. G. Trafton. 1992. Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems. *J. Learn. Sci.* 2, 3 (1992), pp. 277–305.
[41] Philip I. Pavlik and John R. Anderson. 2005. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cogn. Sci.* 29, 4 (2005), 559–586.
[42] Philip I. Pavlik and John R. Anderson. 2008. Using a model to compute the optimal schedule of practice. *J. Exp. Psychol. Appl.* 14, 2 (2008), 101.
[43] André Pereira, Carlos Martinho, Iolanda Leite, and Ana Paiva. 2008. iCat, the chess player: The influence of embodiment in the enjoyment of a game. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, 1253–1256.
[44] Aaron Powers, Sara Kiesler, Susan Fussell, and Cristen Torrey. 2007. Comparing a computer agent with a humanoid robot. In *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI'07)*. IEEE, 145–152.
[45] Aditi Ramachandran, Chien-Ming Huang, and Brian Scassellati. 2017. Give me a break!: Personalized timing strategies to promote learning in robot-child tutoring. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 146–155.
[46] Aditi Ramachandran, Alexandru Litoiu, and Brian Scassellati. 2016. Shaping productive help-seeking behavior during robot-child tutoring interactions. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE Press, 247–254.
[47] G. Saunders. 1988. *Bilingual Children: From Birth to Teens*. ERIC.
[48] Thorsten Schodde, Kirsten Bergmann, and Stefan Kopp. 2017. Adaptive robot language tutoring based on Bayesian knowledge tracing and predictive decision-making. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 128–136.
[49] Sofia Serholt, Christina Anne Basedow, Wolmet Barendregt, and Mohammad Obaid. 2014. Comparing a humanoid tutor to a human tutor delivering an instructional task to children. In *Proceedings of the 14th IEEE-RAS International Conference on Humanoid Robots (Humanoids'14)*. IEEE, 1134–1141.
[50] Daniel Szafir and Bilge Mutlu. 2012. Pay attention!: Designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)*. ACM, New York, NY, 11–20. DOI: https://doi.org/10.1145/2207676.2207679
[51] Fumihide Tanaka, Aaron Cicourel, and Javier R. Movellan. 2007. Socialization between toddlers and robots at an early childhood education center. In *Proceedings of the National Academy of Sciences* 104, 46 (2007), 17954–17958.
[52] Muhammad Tanveer. 2007. Investigation of the factors that cause language anxiety for ESL/EFL learners in learning speaking skills and the influence it casts on communication in the target language. Master's thesis. University of Glasgow, Scotland, UK.
[53] Adriana Tapus, Cristian Tapus, and Maja Mataric. 2009. The role of physical embodiment of a therapist robot for individuals with cognitive impairments. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'09)*. IEEE, 103–107.
[54] Wayne P. Thomas and Virginia P. Collier. 2002. *A National Study of School Effectiveness for Language Minority Students' Long-Term Academic Achievement*. Technical Report. University of California at Santa Cruz, Center for Research on Education, Diversity, and Excellence.
[55] U.S. Census Bureau. 2011. 2010 Census. U.S. Department of Commerce.
[56] Joshua Wainer, David J. Feil-Seifer, Dylan A. Shell, and Maja J. Mataric. 2007. Embodiment and human-robot interaction: A task-based perspective. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'07)*. IEEE, 872–877.
[57] L. R. Welch. 2003. Hidden Markov models and the Baum-Welch algorithm. *IEEE Inf. Theory Soc. Newslett.* 53, 4 (2003), 10–13.
[58] S. M. Williams, D. Nix, and P. Fairweather. 2013. Using speech recognition technology to enhance literacy instruction for emerging readers. In *Proceedings of the 4th International Conference of the Learning Sciences*. 115–120.

Received September 2017; revised June 2018; accepted October 2018